Goals of Geospatial Statistics Workshop and Report

(1) Introduction to spatial and temporal modeling challenges and statistics for spatial data.

(2) Review techniques to assist with spatial and temporal modeling challenges for the EERE modeling teams.

(3) Illustrate examples using common EERE data sets.

(4) Review software for spatial analysis.

(5) Produce a (post workshop) paper with recommended approaches and examples.

Overview

- Why spatial and spatio-temporal statistics?
- How Exploratory Spatial Data Analysis (ESDA) can help interpret spatial data.
- How to use observed data to estimate an unknown process.
- Developing models for point-referenced and areal data.
- How to work with misaligned data and multiple spatial scales.
- Review of Software Options.
- Questions and Discussion.

(1) Introduction to spatial and temporal modeling challenges and statistics for spatial data.

Geospatial Statistics and

Issues in Energy Modeling

Gardar Johannesson and Jeffrey Stewart

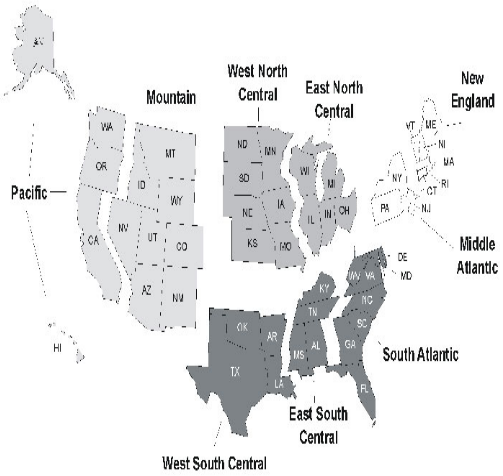
Lawrence Livermore National Laboratory

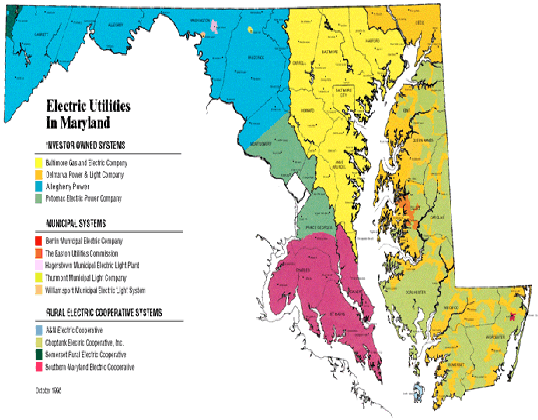
May 10-11, 2005

Common Challenges Working With Spatial and Temporal Data for EERE Modelers

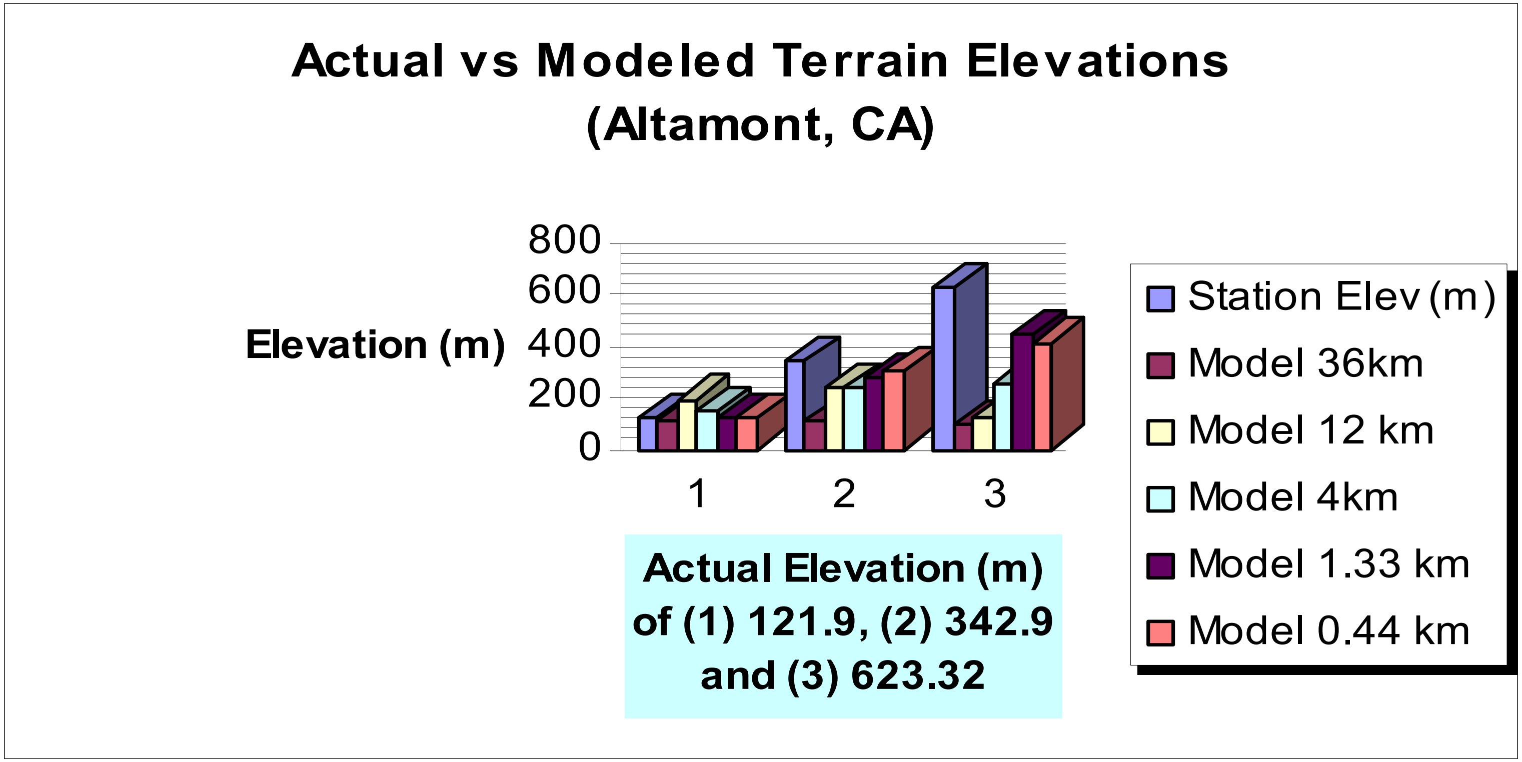
- Aggregating and disaggregating spatial and temporal data between NERC regions, census divisions, numerous utility and county districts.
- Integrating simulated or areal data with sparse point data.
- Incomplete data sets.
- Poor data quality.

Figure 1.3. Electricity Market Model Supply Regions





Physical Models Still Need Improvements Before Relying Solely On Their Results

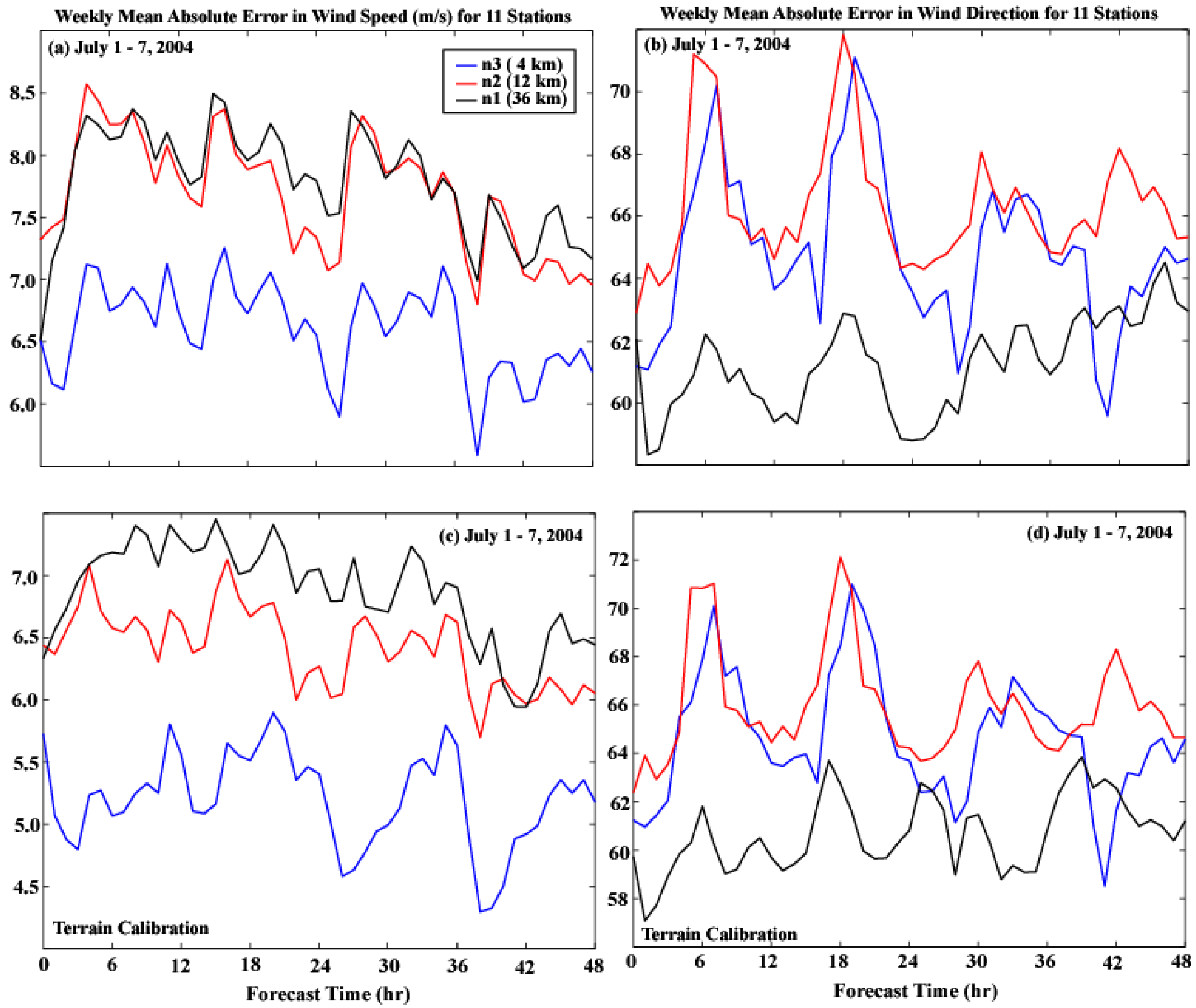


Terrain complexity and other unknown factors contribute to the model errors.

Point data, local physical and regression models can be used to reduce the reliance on single data sources.

Terrain Calibration of Forecast Error Using COAMPS Model and Measurement Station Data

Mean wind speed errors reduced

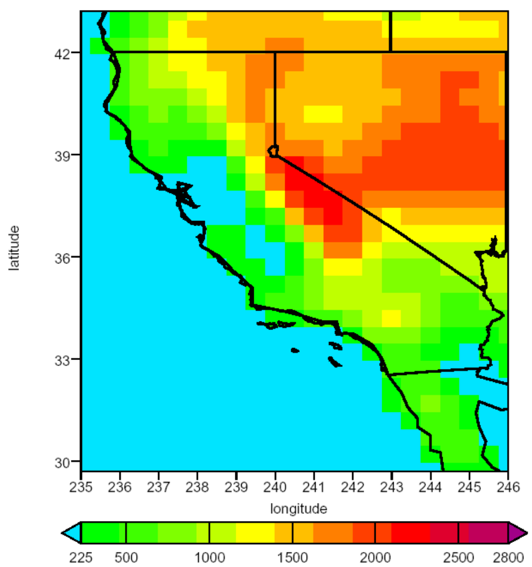


Mean wind direction errors increase

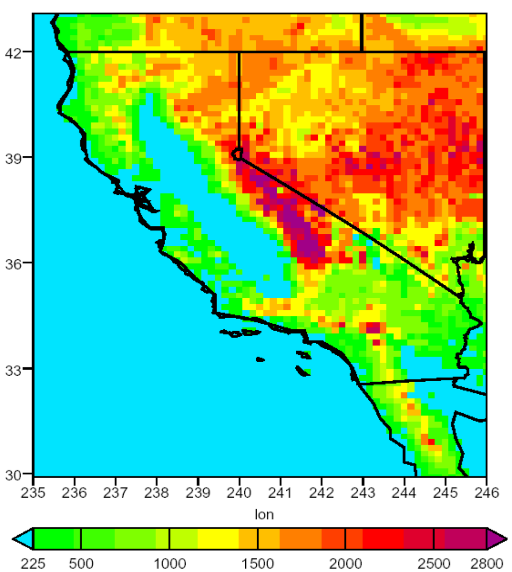
“Data that are close together in space and time are often more alike than those that are far apart”.Cressie

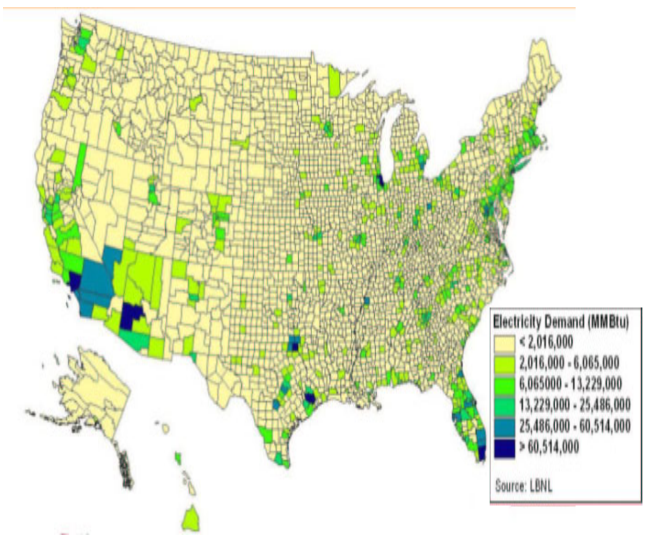
Data representing both physical and socio-economic phenomena tend to demonstrate spatial correlation. The precipitation maps below shows a generally smooth transition The electric demand map shows clustering of high and low energy demand. It indirectly shows the tendency of populations to cluster (see coastal regions).

0.5 deg lat and long



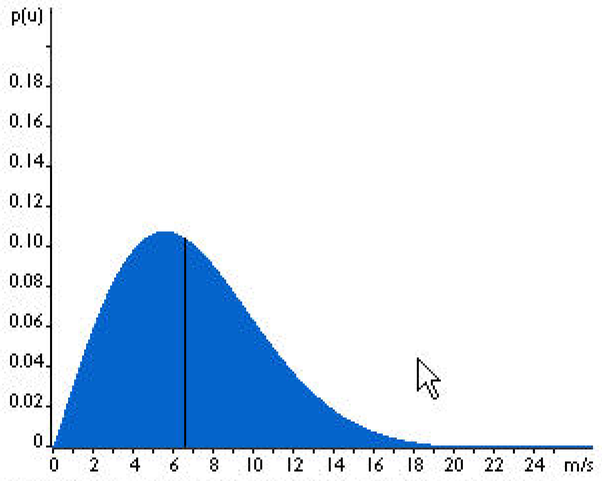
0.1 deg lat and long



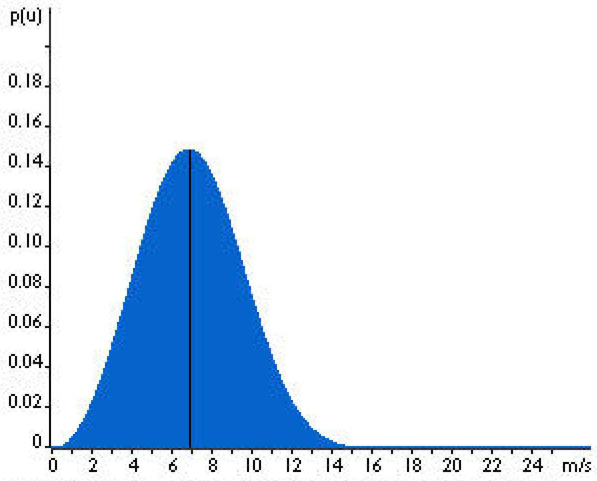


Using Statistical Means Solely Can Lead to Misrepresentation

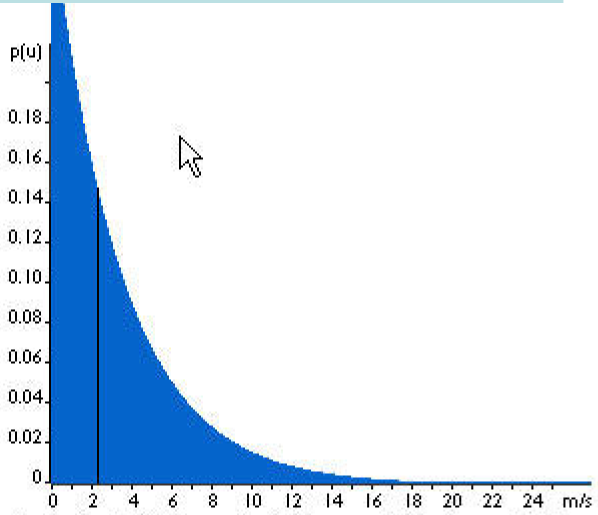
The Mean offers some initial insight into the population or event of concern. Adding a distribution over the population or event improves the ability to estimate or predict. The mean for each one of these graphs is 7 m/s. However, the shape (PDF) of each graph will produce very different wind power estimates. The median wind speeds m/s (6.60, 6.90 and 2.30) begin to give a better indication of the frequency economical winds are available. Regression models using topographical features may allow modelers to estimate improved statistics on a site.



Scale A=7.9; Shape k=2; Mean = 7; Median = 6.60



Scale A=7.84; Shape k=3; Mean = 7; Median = 6.9

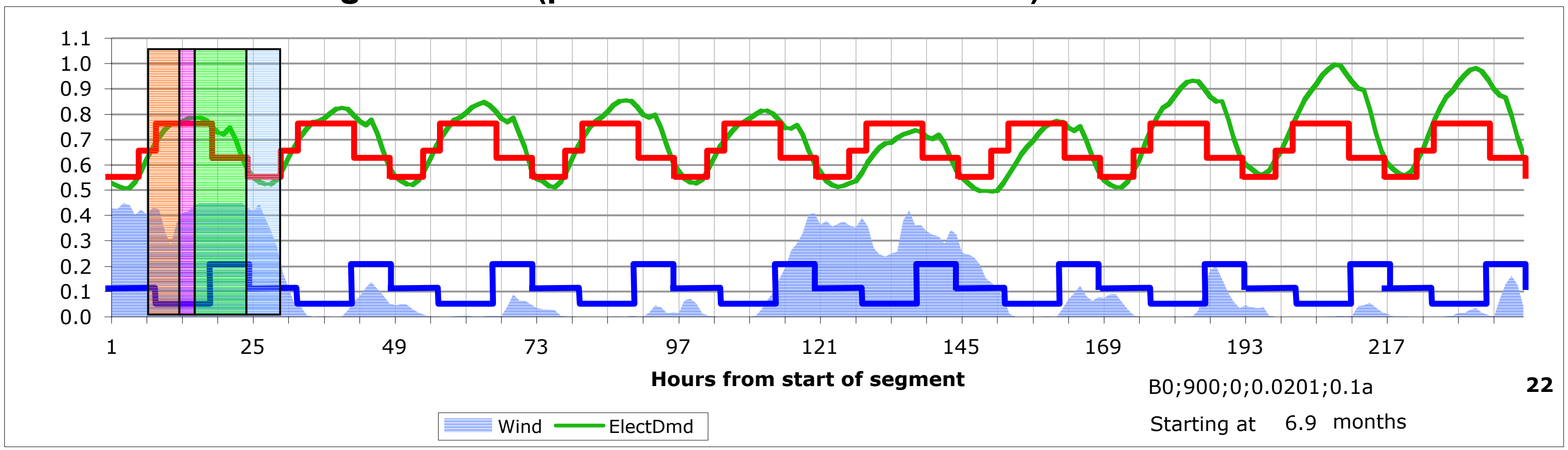


Scale A=3.5; Shape k=1; Mean = 7; Median = 2.30

Some wind turbines do not operate below 3-5m/s or above 25m/s.

Averaging Behavior Over Portions of the Day for Each Season Loses Crucial Information

Demand and wind generation (peak demand scaled to 1kW)



- Averages lose the fact that wind energy comes in bursts, which affect capacities of other generators
- Averages over long periods of the day: don't recognize hourly fluctuations
- Example: averaging loses the fact that there really was no wind generation in the hours that the system demand was at peak at this site.
- Aggregation and averaging is more likely to lead to systematic errors in the penetration of intermittents and changes to the balance of the system

Input Data:Key Take Away

- Data provided by non energy modelers often needs to be modified to appropriately represent the analysis of interest.
- Using data at inappropriate spatial and temporal resolution can lead valid energy models to erroneous results.
- Introducing spatial statistics for data analysis is a convenient and proven way to conduct analysis that can be validated.

General Spatial Model

Statistics relies on various types of stochastic models. The data may be continuous or discrete, spatial aggregations or observations at points in space.The stochastic model is used to summarize existing data or to predict unobserved data. Cressie

The General Spatial Model: $\{Z(s) : s \in D\}$ where:

$s =$ Spatial location belonging to the set D

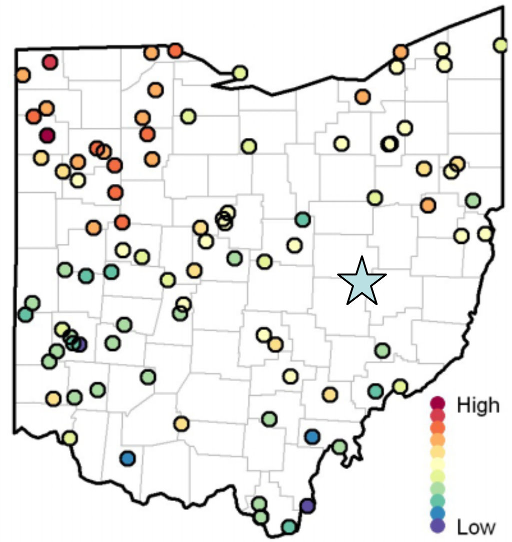
$D =$ Random set

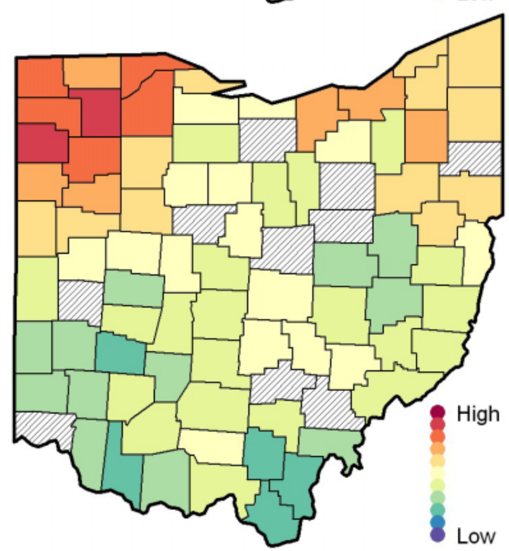
Note: A spatial model may not explain why an event occurs, unlike the more common definition of a model.

What Are We Looking For?

The data we get is not necessarily what we want…

- We might observe point-referenced data, but want a fine-resolution map (interpolation/extrapolation)
- We might observe areal data with missing data for some of the units, but want complete data (prediction/imputation)
- We might observe areal data at a given set of areal units, but need them at a different set of units (misaligned data)





Reason to Use Statistics for Spatial Data

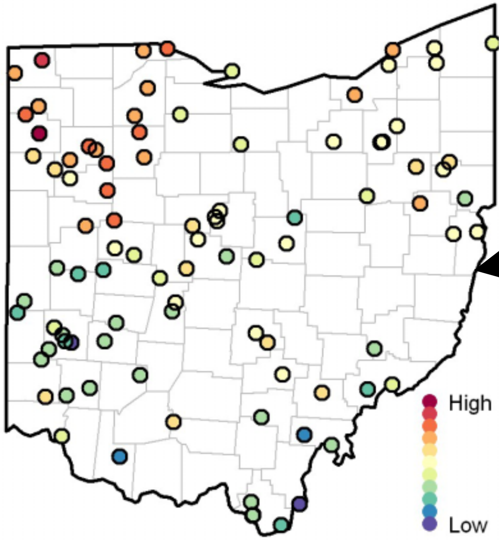
•Make a prediction of a value at a location that was not measured.

How ?

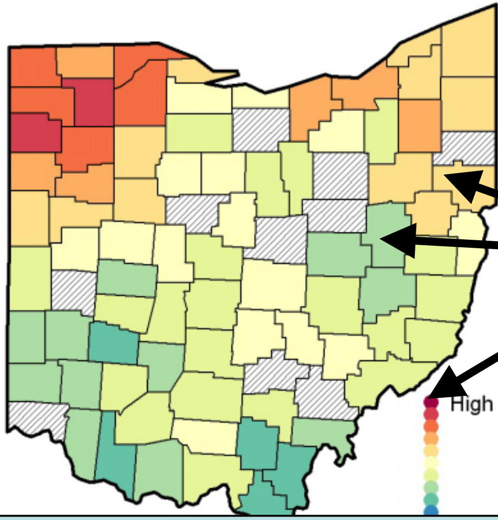
•Obtain data that will allow us to make a probabilistic estimate of the value we want.

•Develop a model of the underlying process and calibrate it from our data.

•Calibrate the model (e.g make a probabilistic estimate of the probability distribution over the error between our model and the true data).



Need to predict un-observed locations



Need to predict missing county data

Our model is basically an estimate of the expected value at any point and is a function of a) the location in space and b) any other relevant "external" data. Make an estimate of the error between the model and the true value. Base this estimate on information from known measurements taken near the location. Make a model of the statistical correlation between points that are near each other.

Data Types

- Basic spatial data-types:
  - Data associated to point locations (point-referenced data) which can come from individual measurement stations (met towers), individual building energy use etc.
  - Data associated to areal units/cells/zones (areal data),examples are summary statistics, simulated data representing statistics on a region, grid, etc.

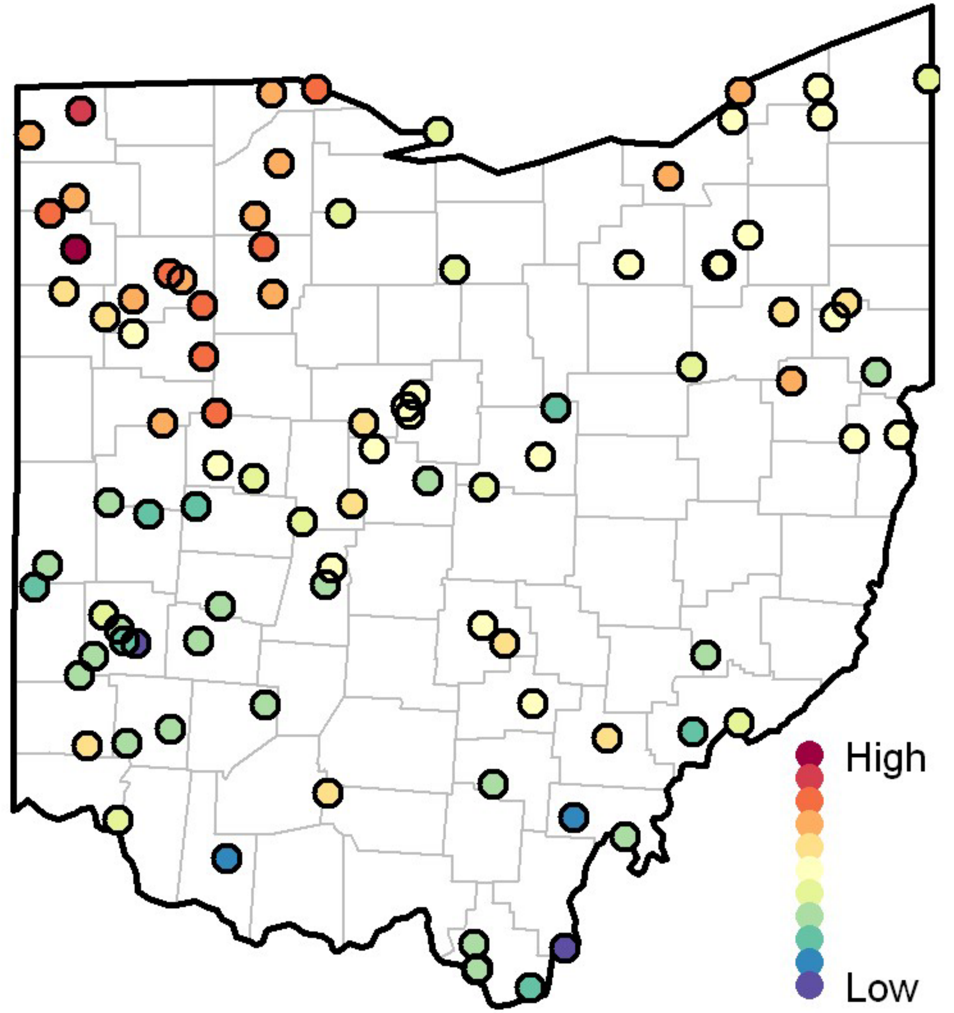
Point-Referenced Data

Each **observation** is associated with a **point location**:

Let $Z(s_1), Z(s_2), \ldots, Z(s_n)$ denote *n* observations associated with the spatial point-locations $s_1, s_2, \ldots, s_n$

Example: Observed average wind-speed in a given time period at given sites

The data can be observed at spatially irregular sites or on a regular grid



Observed data at 92 spatially irregular sites

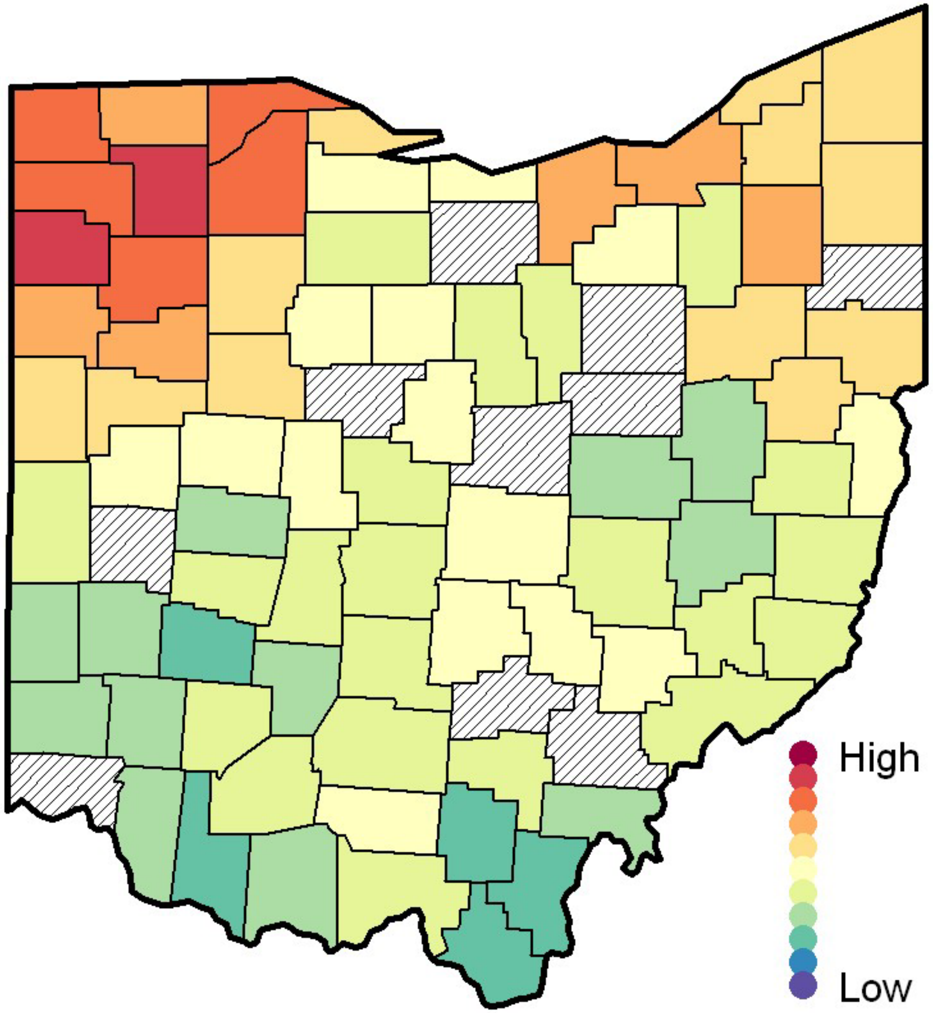
Areal Data

Each **observation** is associated with an **areal unit** (**cell, zone**)

Let $Z(D_1), Z(D_2), \ldots, Z(D_n)$ denote n observations associated with the areal units $D_1, D_2, \ldots, D_n$

Two types of areal data:

- **Areal aggregation**. The observed data is generated by an aggregation process (e.g., counties' average wind speed)
- **Areal explicit**. The data is explicit to the areal units (e.g., counties' budget)

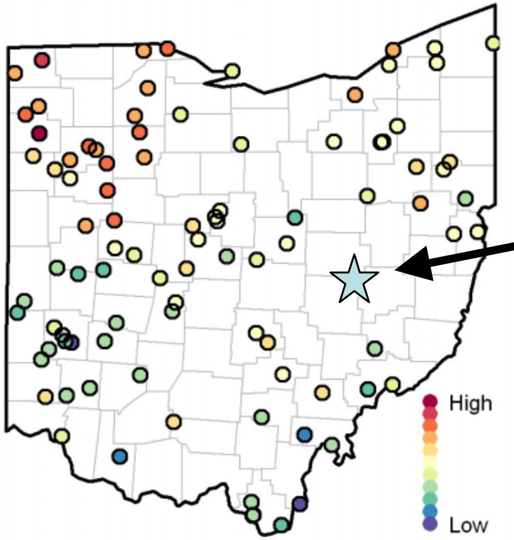


County-aggregated data; some counties have missing data

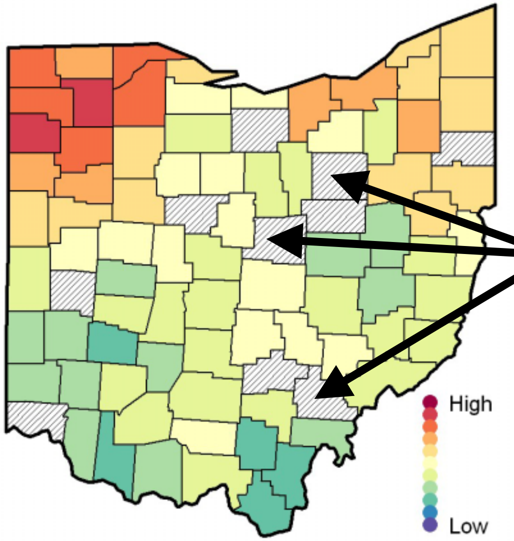
What Are We Looking For?

The data we get is not necessarily what we want…

- We might observe point-referenced data, but want a fine-resolution map (interpolation/extrapolation)
- We might observe areal data with missing data for some of the units, but want complete data (prediction/imputation)
- We might observe areal data at a given set of areal units, but need them at a different set of units (misaligned data)



Need to predict un-observed locations



Need to predict missing county data

Defining The Unknown Process

- The Unknown (Target) Process is the data we are trying to estimate to explain a process, event, phenomena. We may not need to understand "why" it happened or is expected to happen, but we do model when it should have happened or may happen.

The Unknown Process

Make a formal separation between what is observed and what is sought after

- Point Data:

$$Z(s_i) = Y(s_i) + \varepsilon_{p,i}\ ;\ \ i = 1,\ \ldots,\ n_p$$

- Areal Data:

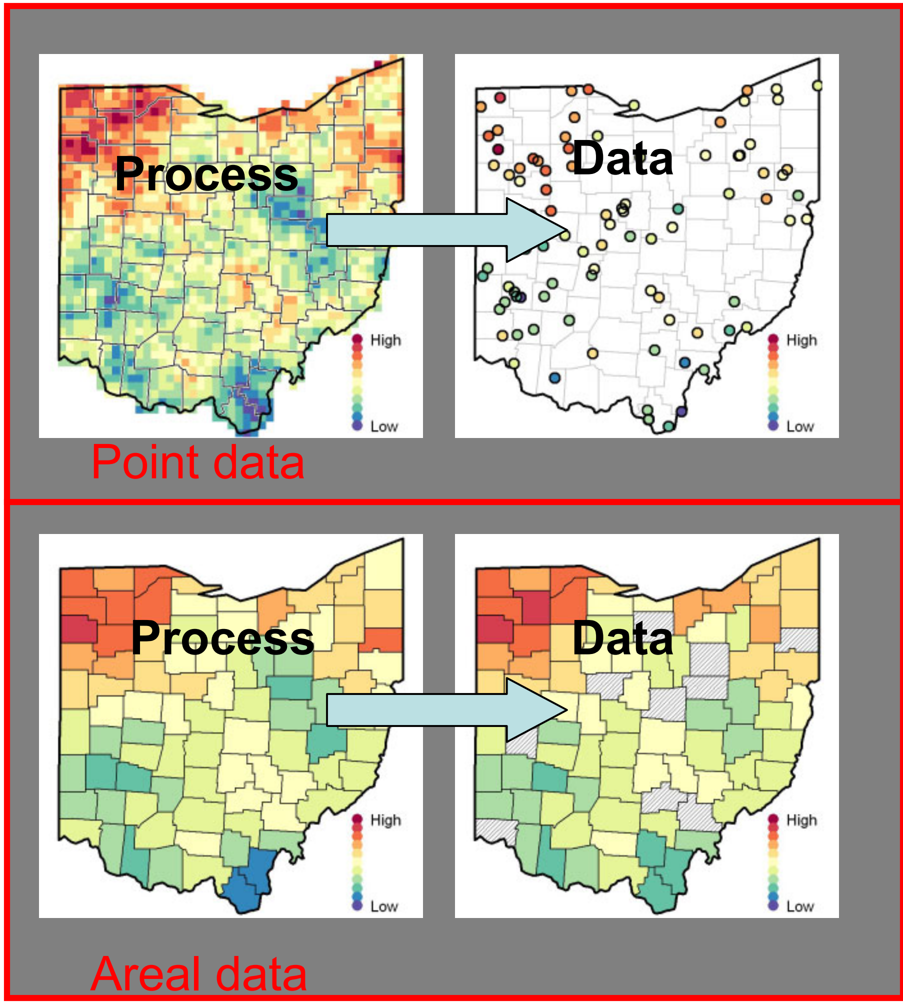
$$Z(D_j) = Y(D_j) + \varepsilon_{a,j}\ ;\ \ j = 1,\ \ldots,\ n_a$$

Where we denote by:

Y = the true, unknown process, the Y-process

ε = the data-error (if any)

The data-error might be zero for some data



Multiple Data Types

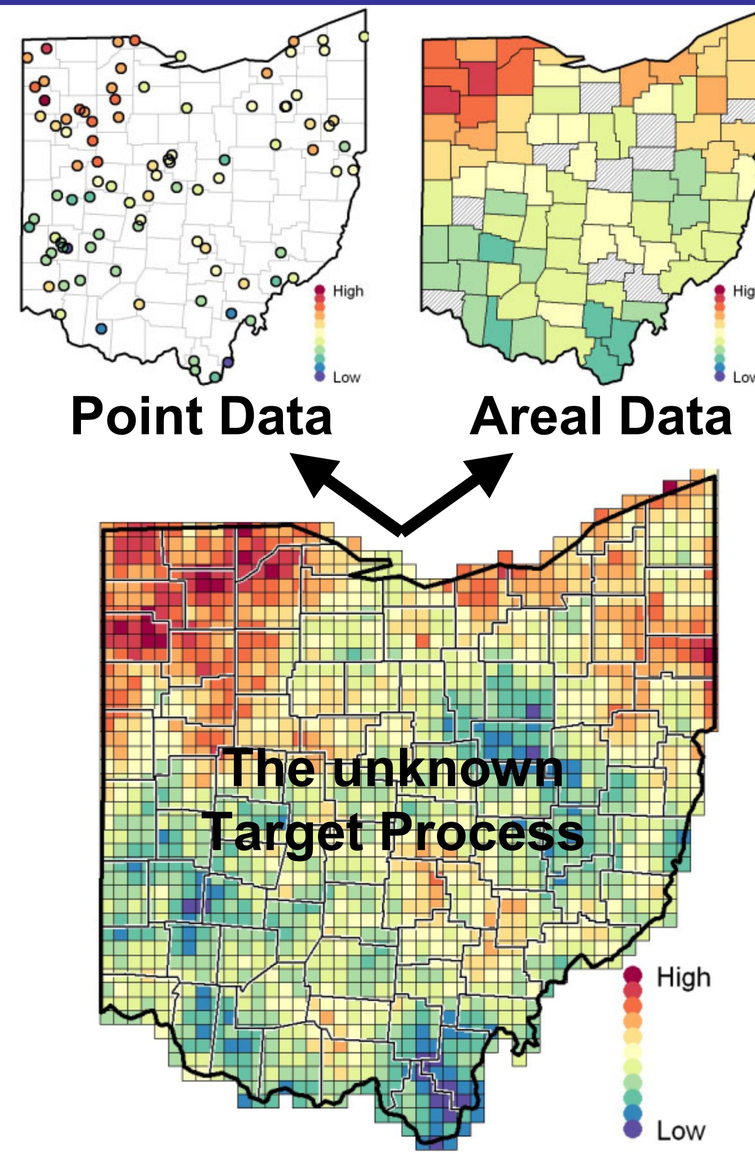
The same Y-process can often be the generating mechanism for both point-data and areal-data:

- $Z(s) = Y(s) + \varepsilon_p$
- $Z(D) = Y(D) + \varepsilon_a$, where

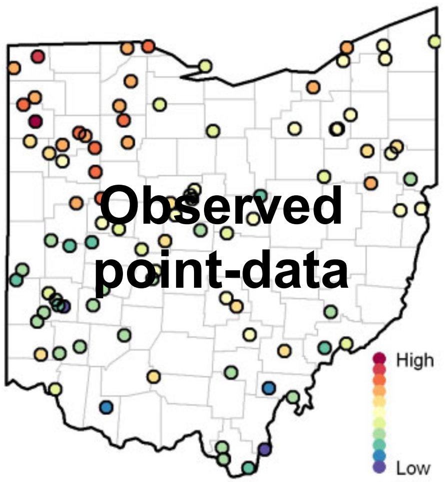
$$\boxed{Y(D) = \text{avg}\{\ Y(s) : s \text{ in } D\ \}}$$

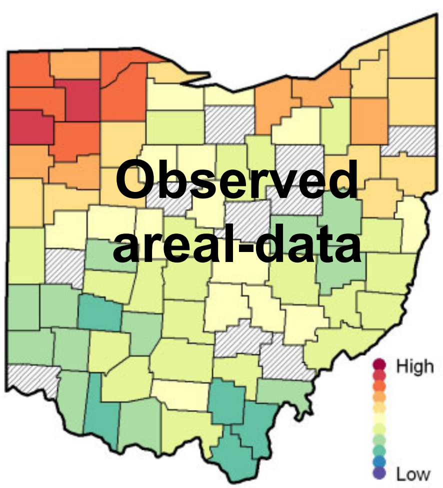
Hence, the value of the Y-process in areal unit D is just given by the average (integration) of Y over D

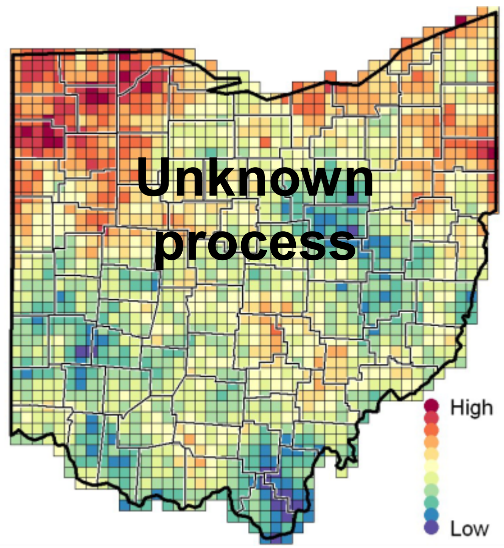
There are other cases where the Y-process is only well defined on a given set of areal units, and aggregation of those units

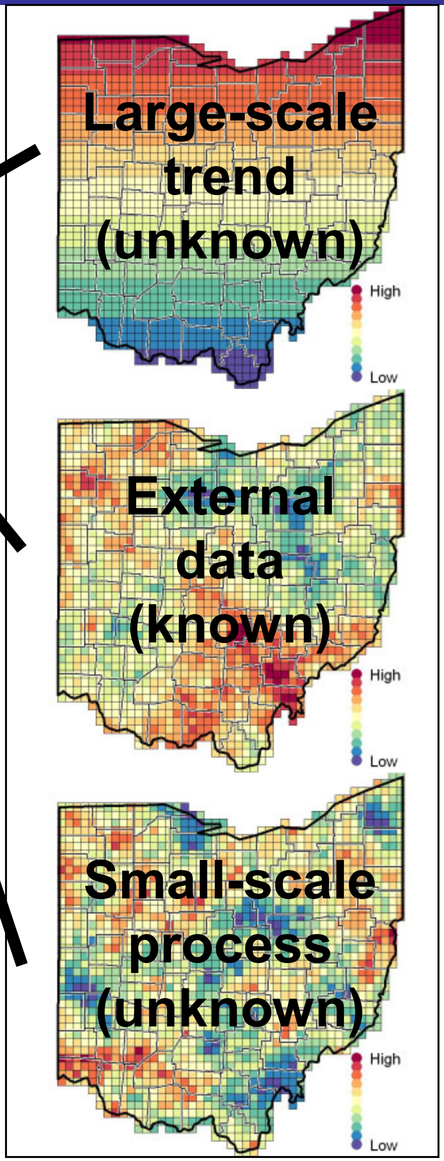


Modeling the Unknown Process









Break down the variation in the unknown target process into contributing components and then analyze and model each component

- ✓ Large-scale variation (trend)
- ✓ External explanatory data
- ✓ Spatially correlated small-scale variation

Geospatial Statistics and Issues in Energy Modeling

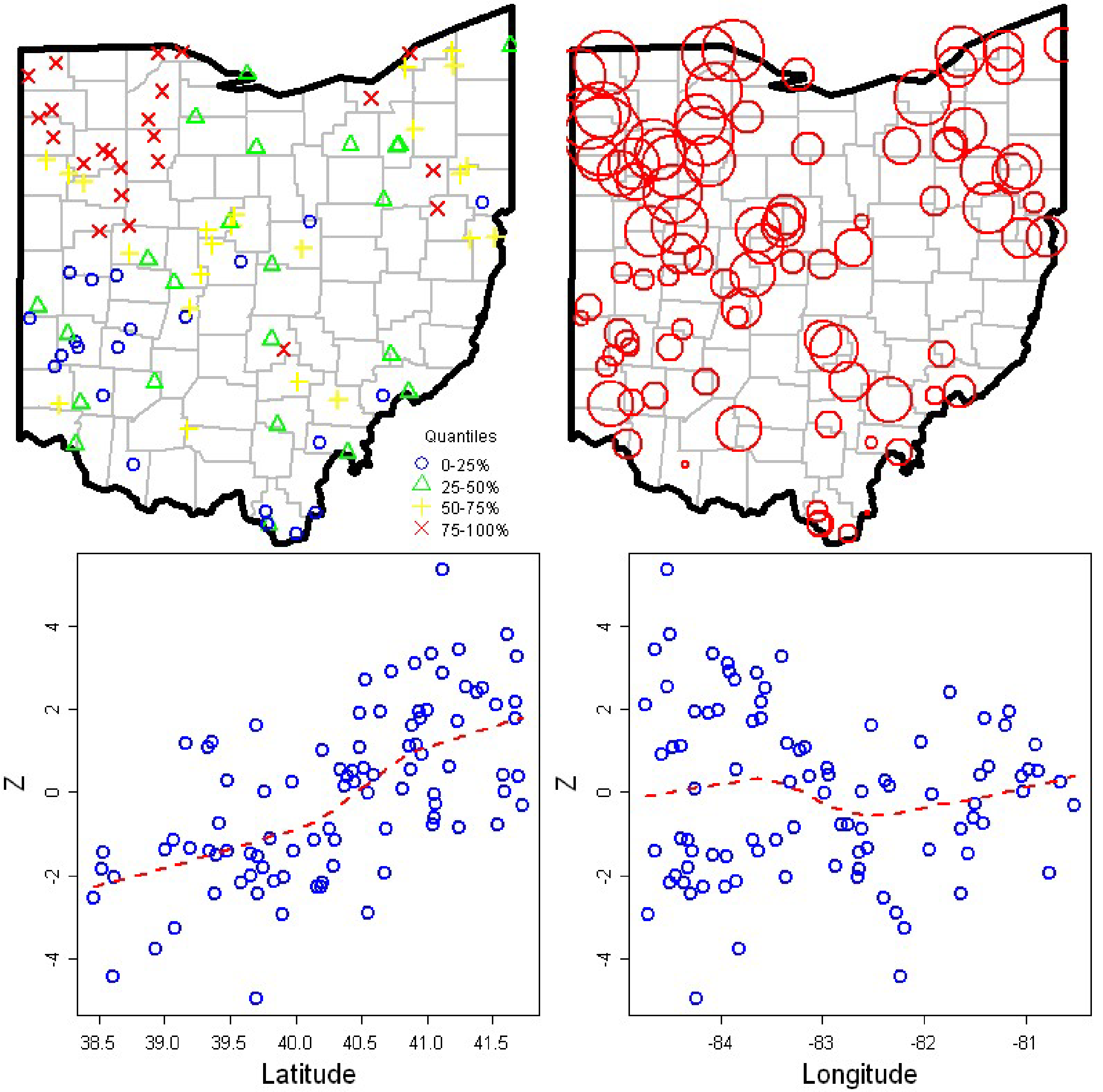
Gardar Johannesson and Jeffrey Stewart
Lawrence Livermore National Laboratory
May 10-11, 2005

Defining Exploratory Spatial Data Analysis

- ESDA should always be the first step in any data analysis
- ESDA is mostly graphics-based approach to explore the basic characteristics of the observed data
- In ESDA, we particularly look for approaches to explore:
  - The possibility of large-scale spatial trend in the data and how that trend can be captured/model.
  - The (marginal) variation in the data; the mean, the spread of the data, etc.
  - Possible correlation to external readily-available data and how that data can be used to help in explaining (and reduce) the variation in the observed data.
  - The spatial correlation in the data and how it varies with distance.

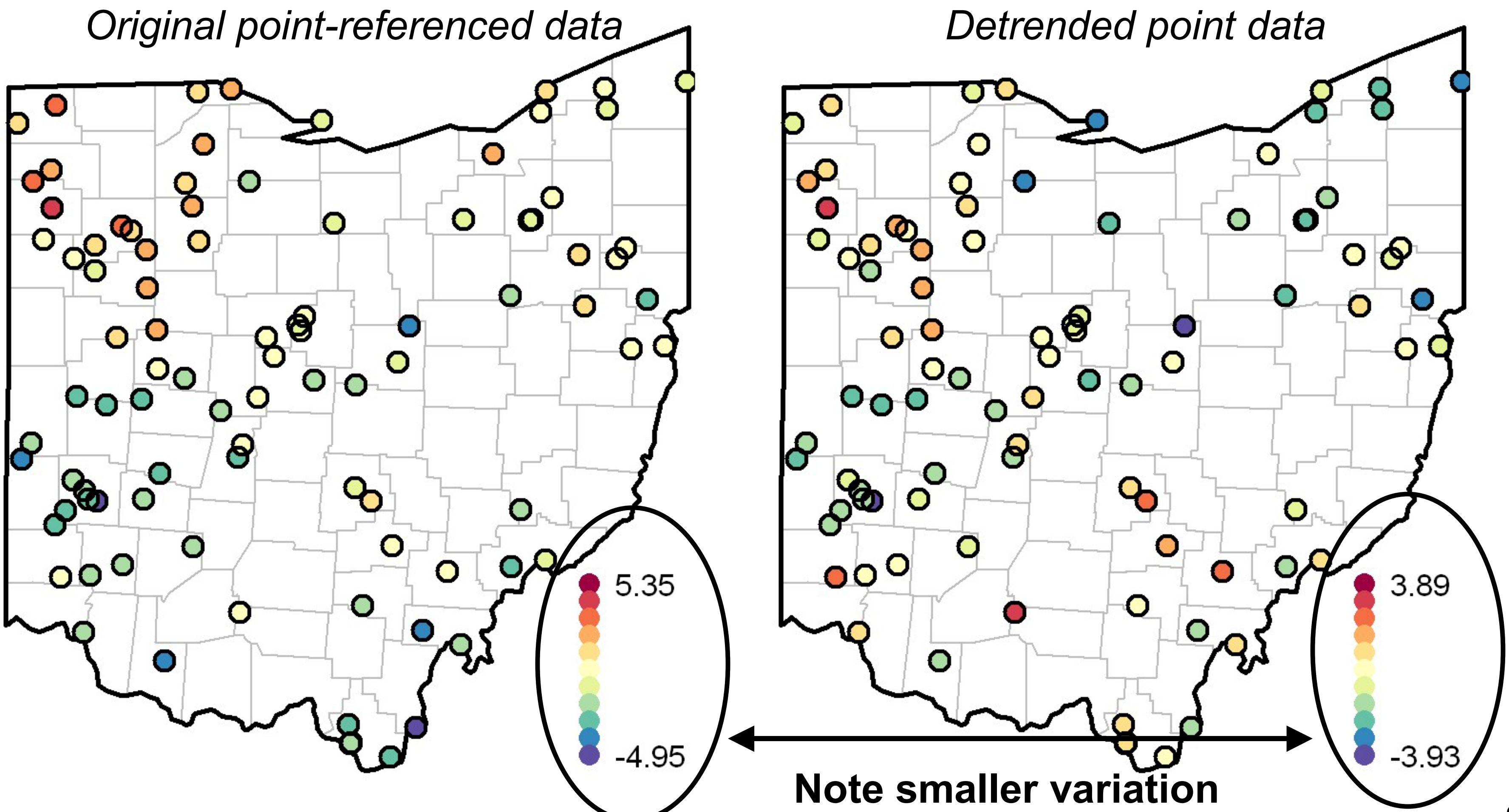
Observing Spatial Trend

- Visually easy to see large scale-trends in areal data
- There are numerous graphical ways to visualize potential trend in point-referenced data
  - Label data-points according to rank (quartiles)
  - Use plot symbols that reflect the observed value at each point
  - Marginal East/North scatter plots



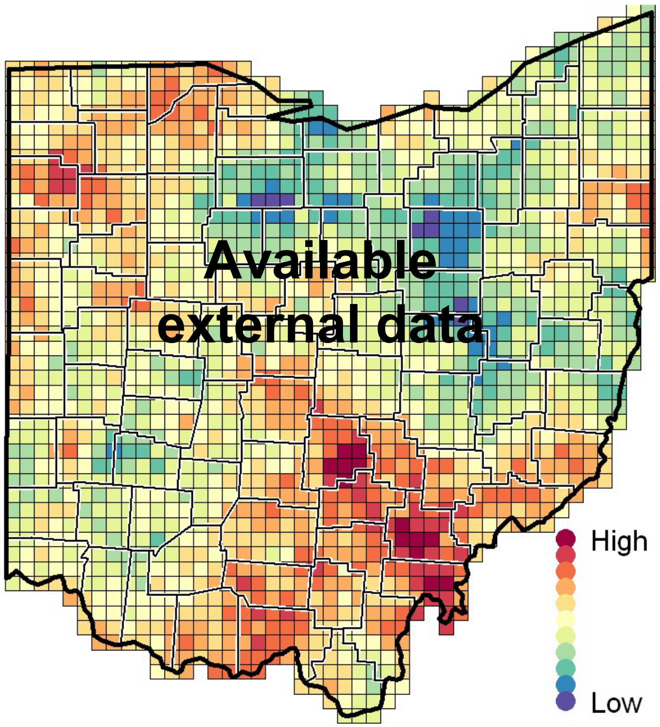
Detrending Point-Data Reduces the Large Scale Variation

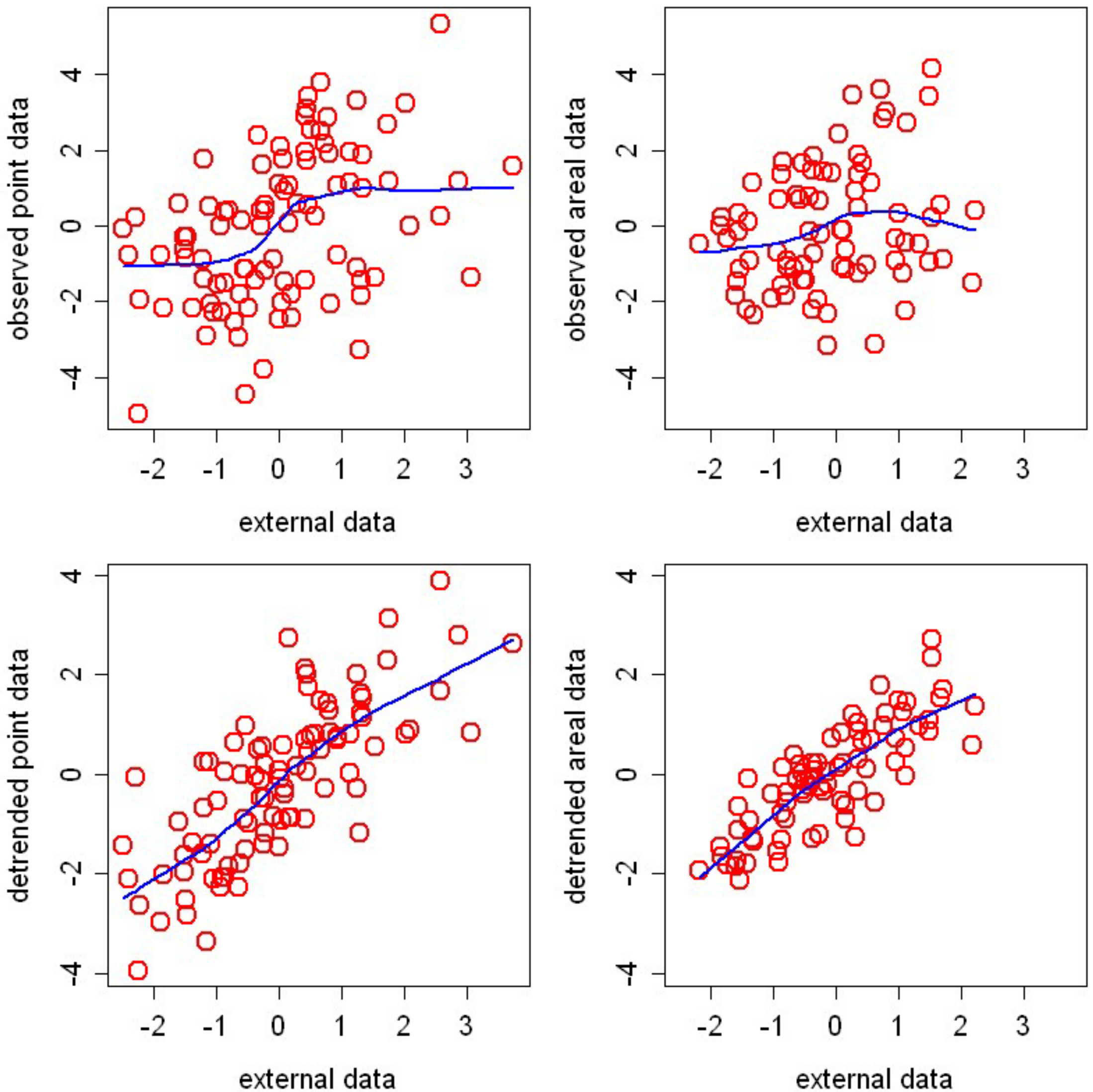
The large-scale trend is subtracted from the point data, yielding **detrended point-data** that has smaller variation



Using External Explanatory Data

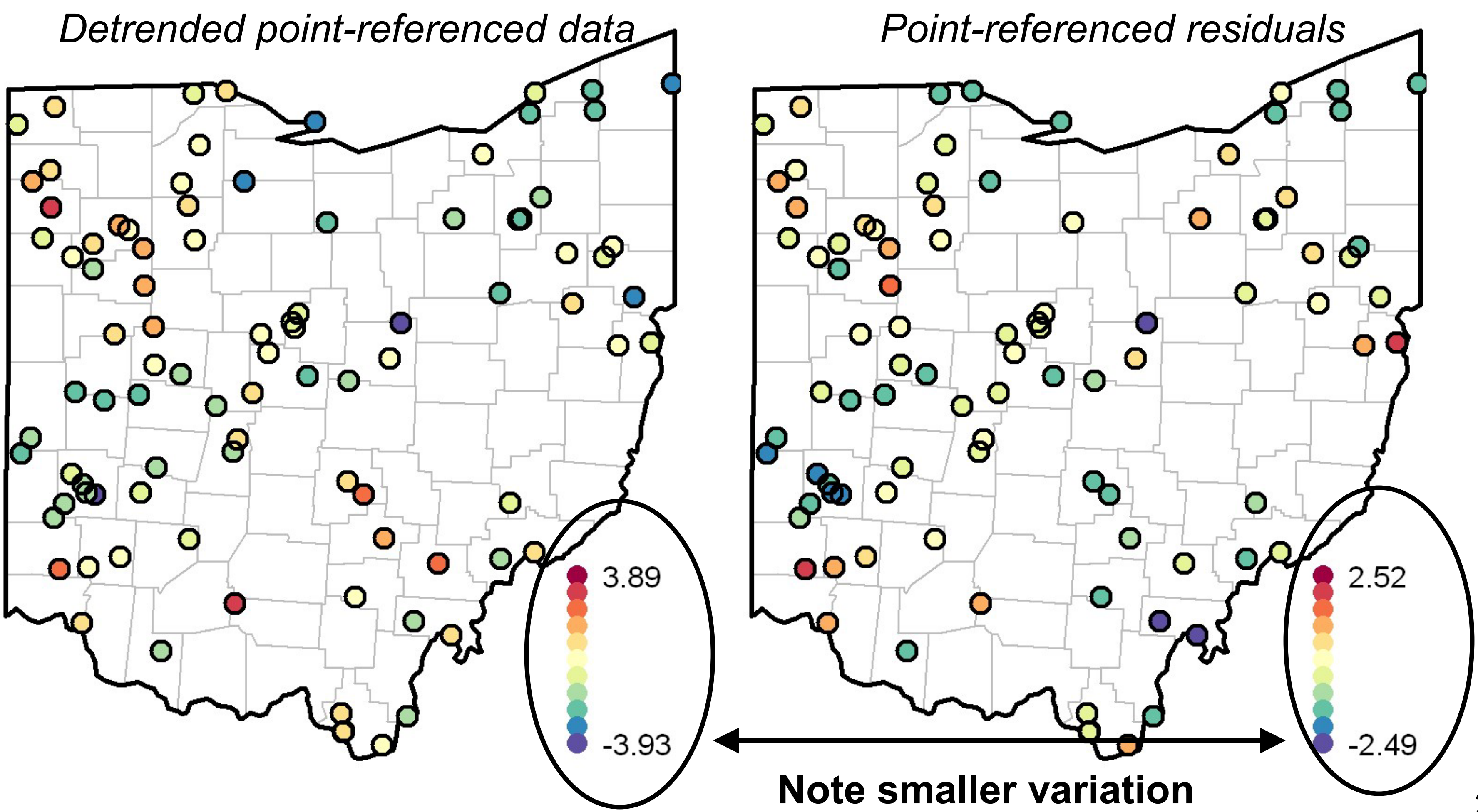
There might be available external data that is correlated with the observed data, and therefore useful for prediction





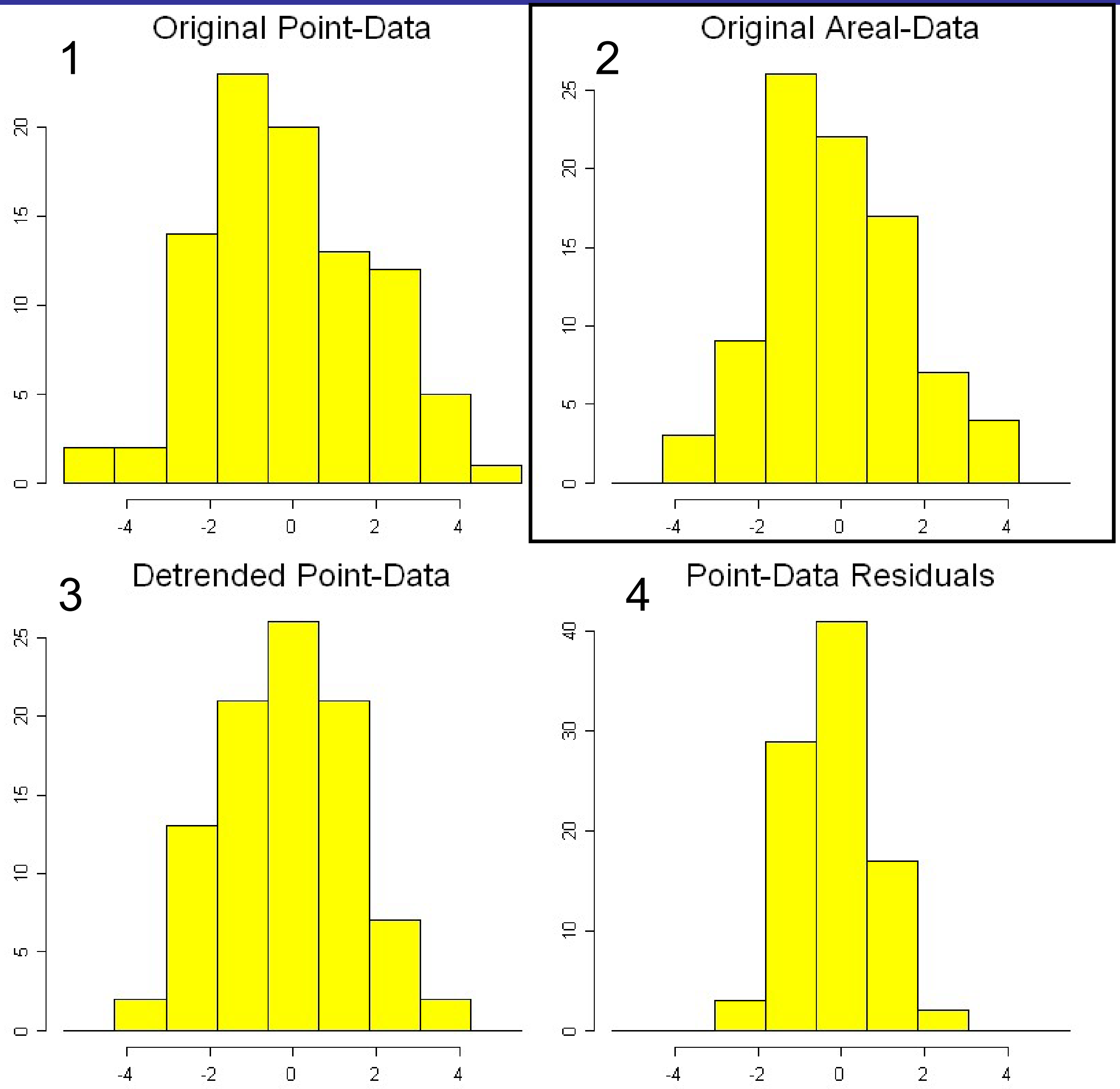
Additional Analysis Using Point-Data Residuals

The contribution of the external data is subtracted from the detrended point-data, yielding **point-data residuals** that have even smaller variation



Variation Reduction

- Variation in data can be explored through histograms
- Aggregation reduces the variation (see histogram 1 and 2 of original point-data vs. original areal-data)
- Detrending reduces the variation (see 1 and 3)
- The use of explanatory variables reduces the variation even further (see 3 and 4)



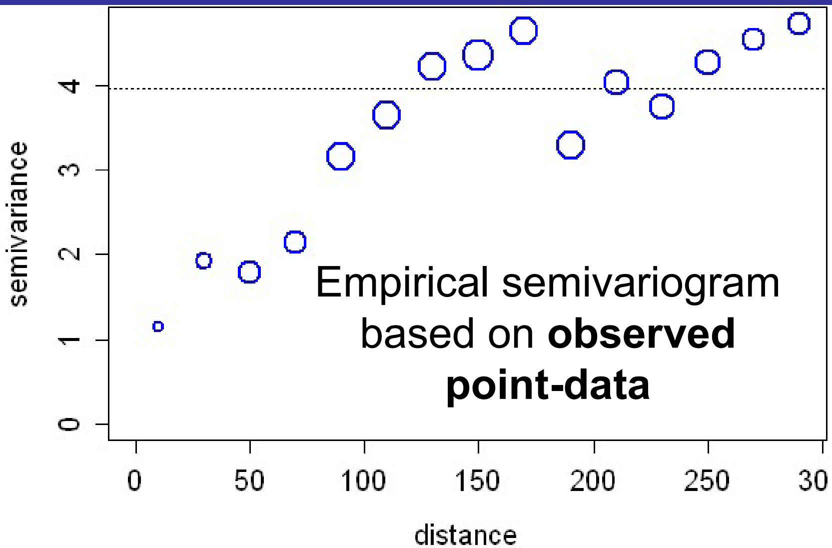
Spatial Association (Point-Data)

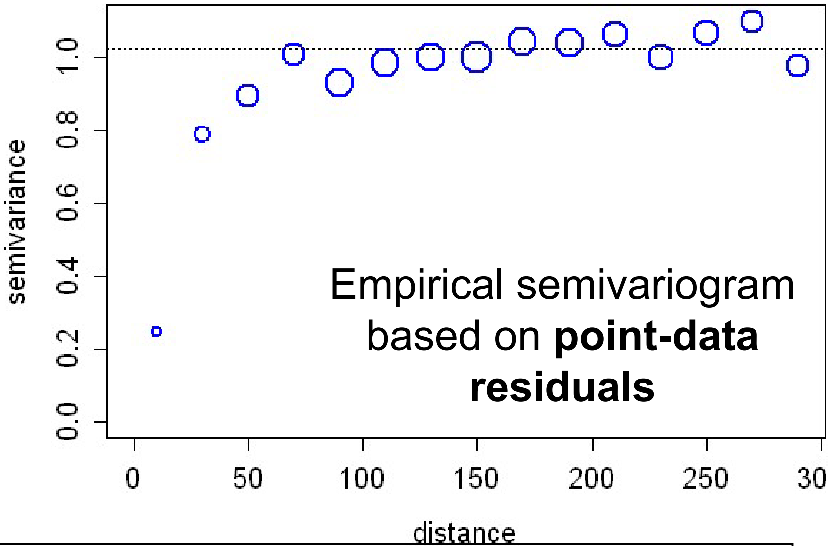
Point-Referenced Data: The goal is to gain information about the spatial association in point-referenced data

The empirical semivariogram: Provides information about the spatial association in point-referenced data as a function of distance between points

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{\substack{(s_i, s_j) \\ \text{in } N(h)}} \left(Z(s_i) - Z(s_j)\right)^2$$

Where h is a given distance and $N(h)$ is the set of points (approximately) separated by h

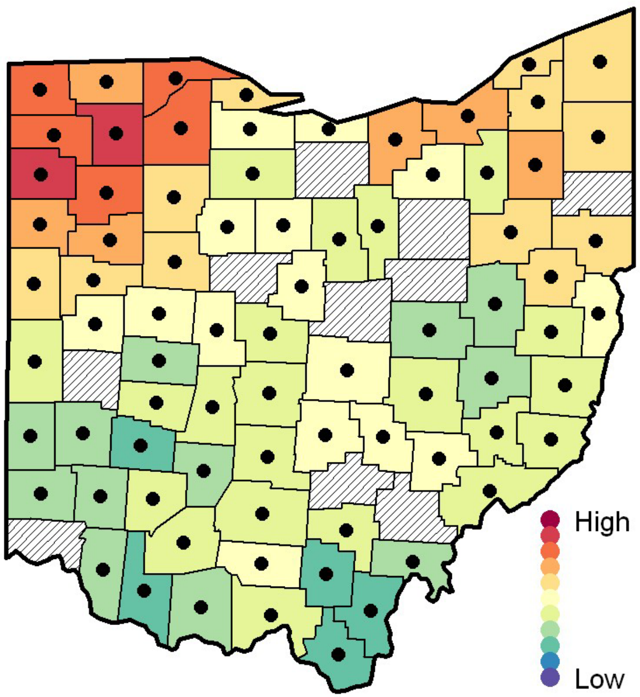


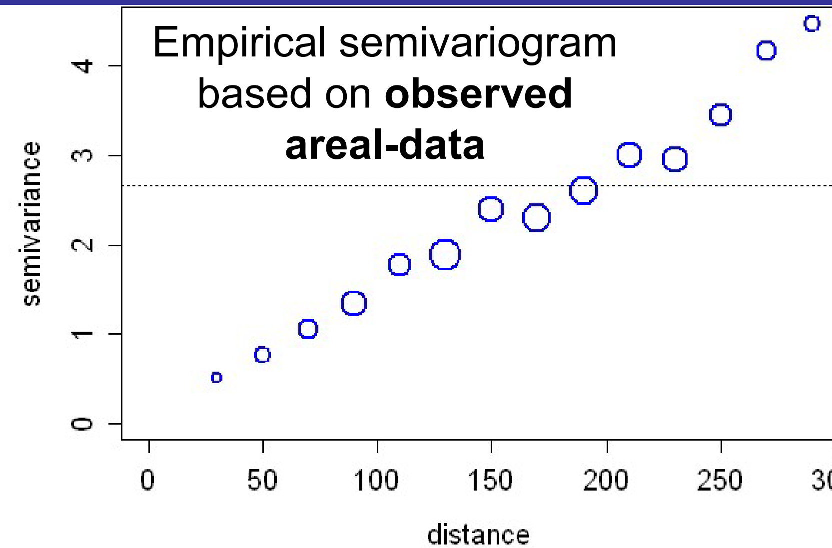


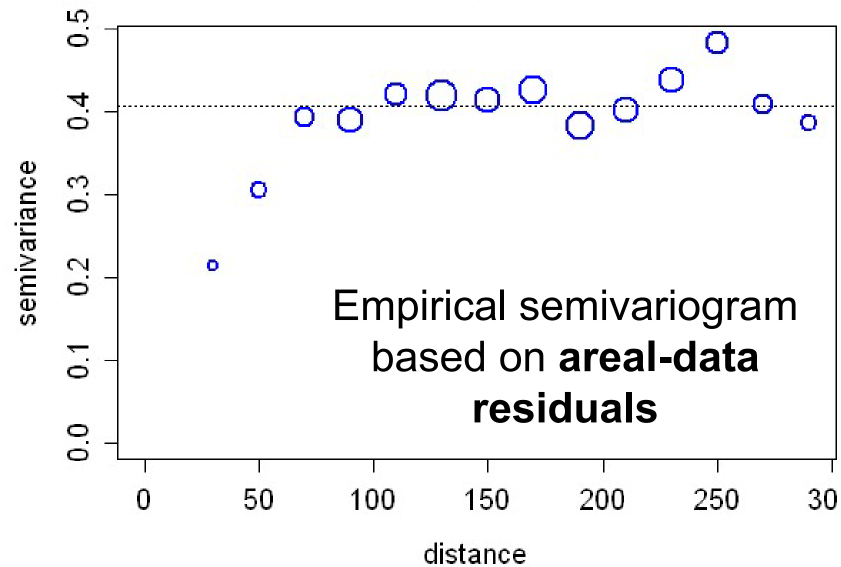
Spatial Correlation is strongest near zero and does not exist above dashed line

Spatial Association (Areal-Data)

One can use the empirical semivariogram to explore the spatial association in **areal data**. That requires a distance measure between areal units, which can be taken as the distance between the units' centroids







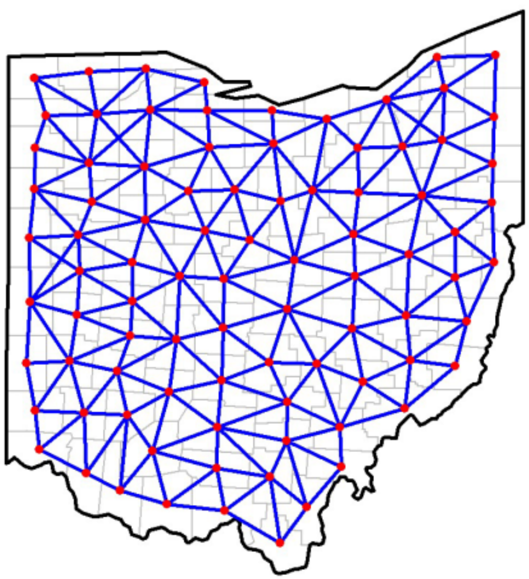
Spatial Proximity of Areal Units

An alternative to use areal units' centroids to define a distance-measure between units is to **define the spatial proximity through a neighbor relationship**

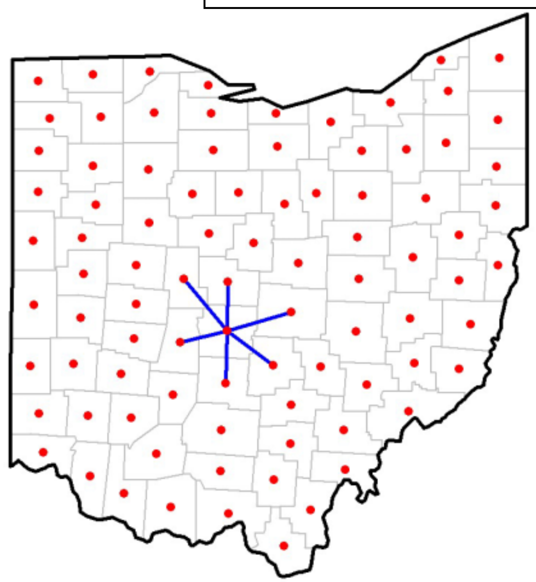
The **proximity matrix** $W = (w_{ij})$:

$$w_{ij} = \begin{cases} 1 & \text{if } D_i \sim D_j \\ 0 & \text{otherwise} \end{cases}$$

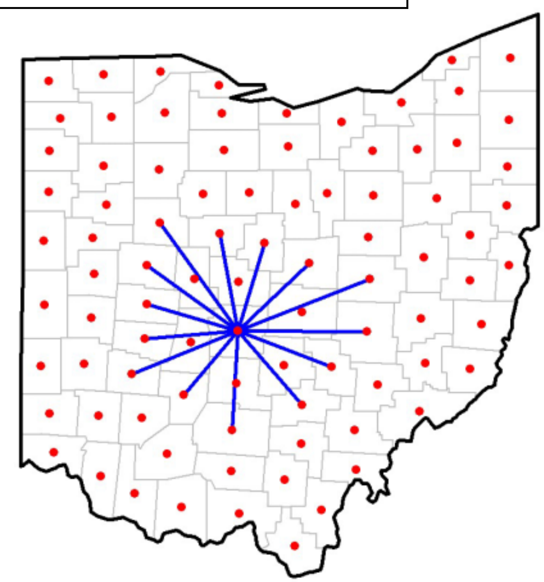
$D_i \sim D_j$: D_i and D_j are neighbors



Counties that share borders are neighbors



First order neighbors of a given county



Second order neighbors of a given county

Spatial Association via Proximity

The goal is to measure (empirically) the spatial association in areal data

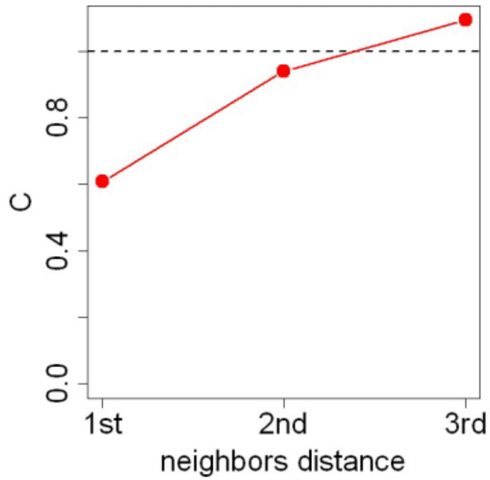
Geary's C provides information about the spatial association in areal data by using the spatial proximity matrix W

$$C = \frac{\sum_{i,j} w_{ij} \left(Z(D_i) - Z(D_j) \right)^2 \Big/ \sum_{i,j} w_{ij}}{\sum_i \left(Z(D_i) - \bar{Z} \right)^2 \Big/ (n-1)}$$

Geary's C has similar interpretation as the semivariogram

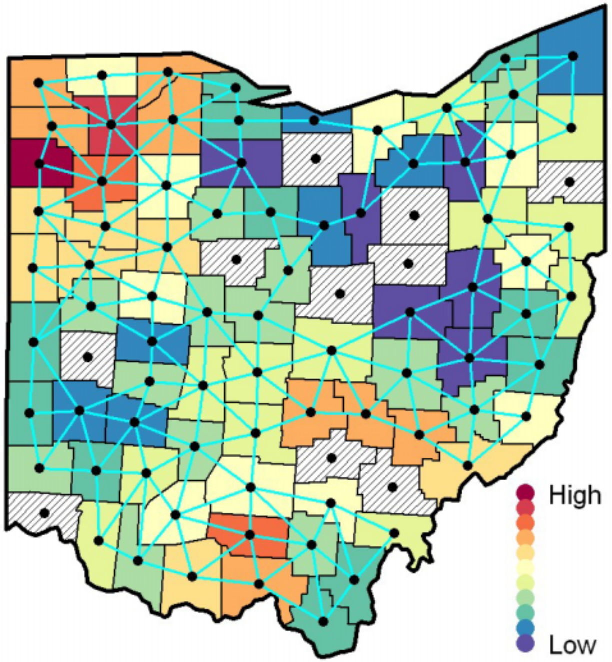
The graph to the left shows Geary's C computed for 1st, 2nd, and 3rd order spatial neighbors using areal-data residuals

There is very little spatial association outside 1st order neighbors



ESDA: Geary's C and Moran's I

Geary's C and Moran's I statistics provides information about the spatial associating in data defined on areal units



$$C = \frac{\sum_{i,j} w_{ij}\left(Z(D_i) - Z(D_j)\right)^2 \Big/ \sum_{i,j} w_{ij}}{\sum_i \left(Z(D_i) - \bar{Z}\right)^2 \Big/ (n-1)}$$

$$I = \frac{\sum_{i,j} w_{ij}(Z(D_i) - \bar{Z})(Z(D_j) - \bar{Z}) \Big/ \sum_{i,j} w_{ij}}{\sum_i \left(Z(D_i) - \bar{Z}\right)^2 \Big/ n}$$

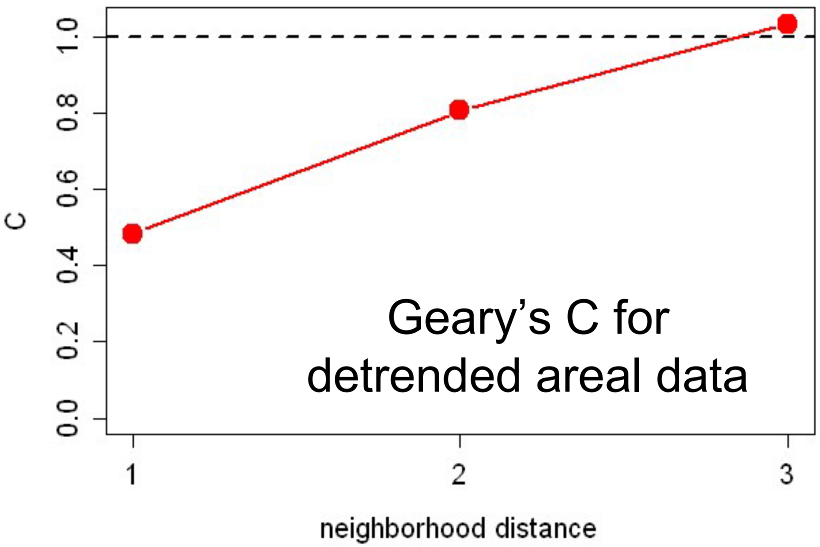
Where recall that $W = (w_{ij})$ is the proximity matrix. Under no spatial association, the expected value of C and I is 1 and -1/(n-1), respectively. Low value for C and a value of I distant from -1/(n-1) hint at spatial association

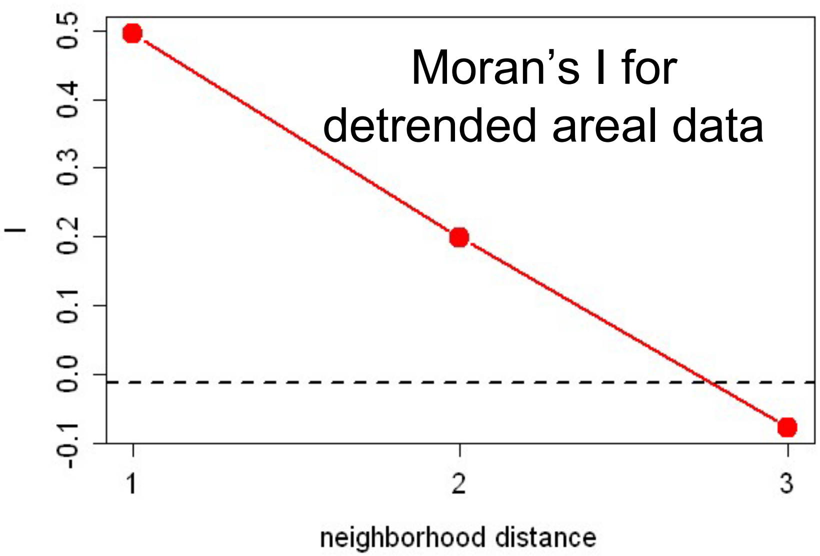
ESDA: Geary's C and Moran's I

Example of Geary's C and Moran's I applied to areal county data (from previous slide) for 1st, 2nd, and 3rd order neighbors

The two provide the same information; 3rd order neighbors show little or no association

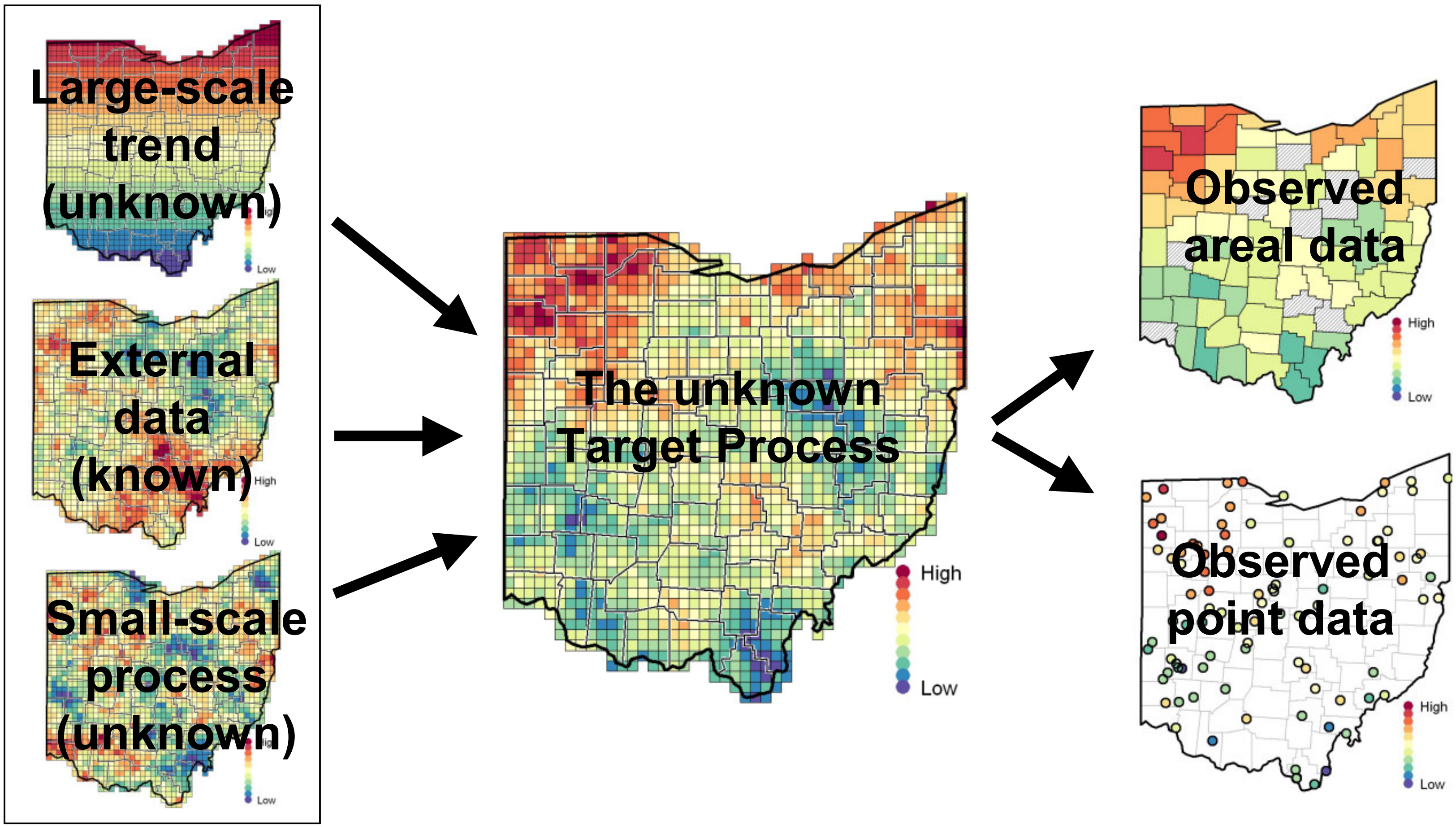
It is possible to perform a formal test of significance for both Geary's and Moran's association statistics, but they are mostly used as exploratory tools





Spatial Modeling Recap

The goal is to construct a model that relates the observed spatial data to the unknown process of interest and can be used to predict the process at any site/area of interest, along with an uncertainty measure



Models for Point-Referenced Data

"Classical" Geostatistical Model

- **The Data Model**:

  $$Z(s_i) = Y(s_i) + \varepsilon(s_i)\ ;\ \ i = 1, \ldots, n$$

  where:

  $Y(s_i)$ = unknown process

  $\varepsilon(s_i)$ = observation errors

- **The Process Model**:

  $$Y(s_i) = \mu(s_i) + f(x_{\mathrm{i}}) + \delta(s_i)$$

  where:

  $\mu(s_i)$ = large-scale trend

  $f(x_{\mathrm{i}})$ = external predictor

  $\delta(s_i)$ = small-scale variation

It then remains to specify the various terms of the model

- **The error process $\varepsilon(s_i)$:**

  Assumed independent with zero mean and variance σ^2

- **The large-scale trend $\mu(s_i)$:**

  Assumed deterministic, for example $\mu(s_i) = a + b\ \mathrm{lat}(s_i)$

- **The external predictor $f(x_{\mathrm{i}})$:**

  A function of external data x_i

- **The small-scale process $\delta(s_i)$:**

  Assumed to be zero mean and spatially correlated process:

  $$\mathrm{Cov}(\delta(s_i),\ \delta(s_j)) = \tau^2\ K(|s_i - s_j|)$$

  where τ^2 is the variance and K is a distance based correlation function

Modeling the Spatial Correlation

- The distance-based correlation function $K(|s_i - s_j|)$ of the small-scale process needs to be specified
- Some options are:

  Exponential: $\exp(-d/\phi)$

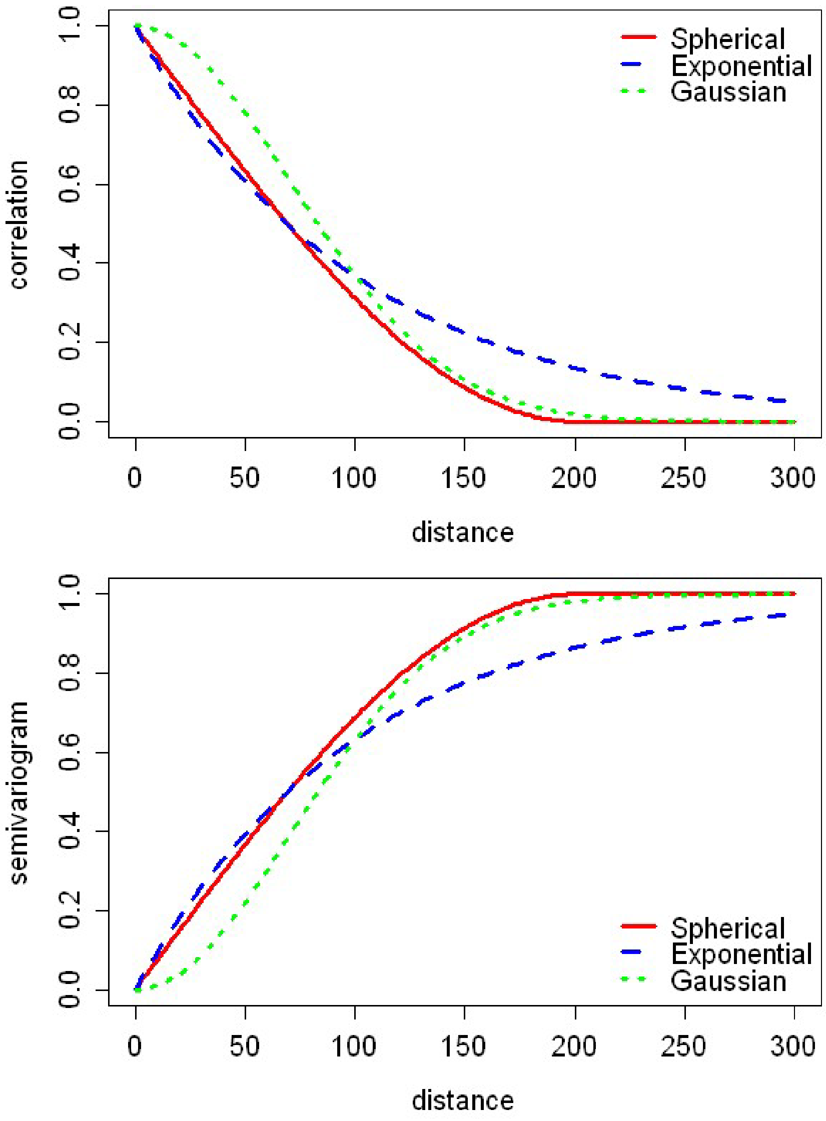
  Gaussian: $\exp(-(d/\phi)^2)$

  where d is distance and ϕ is a range parameter, controlling the extend of the spatial correlation
- The **variogram**: It is the case that

$$\mathrm{Var}(\delta(s_i) - \delta(s_j)) = 2\left[\, \tau^2 (1 - K(|s_i - s_j|))\, \right]$$

  Define the **semivariogram** as:

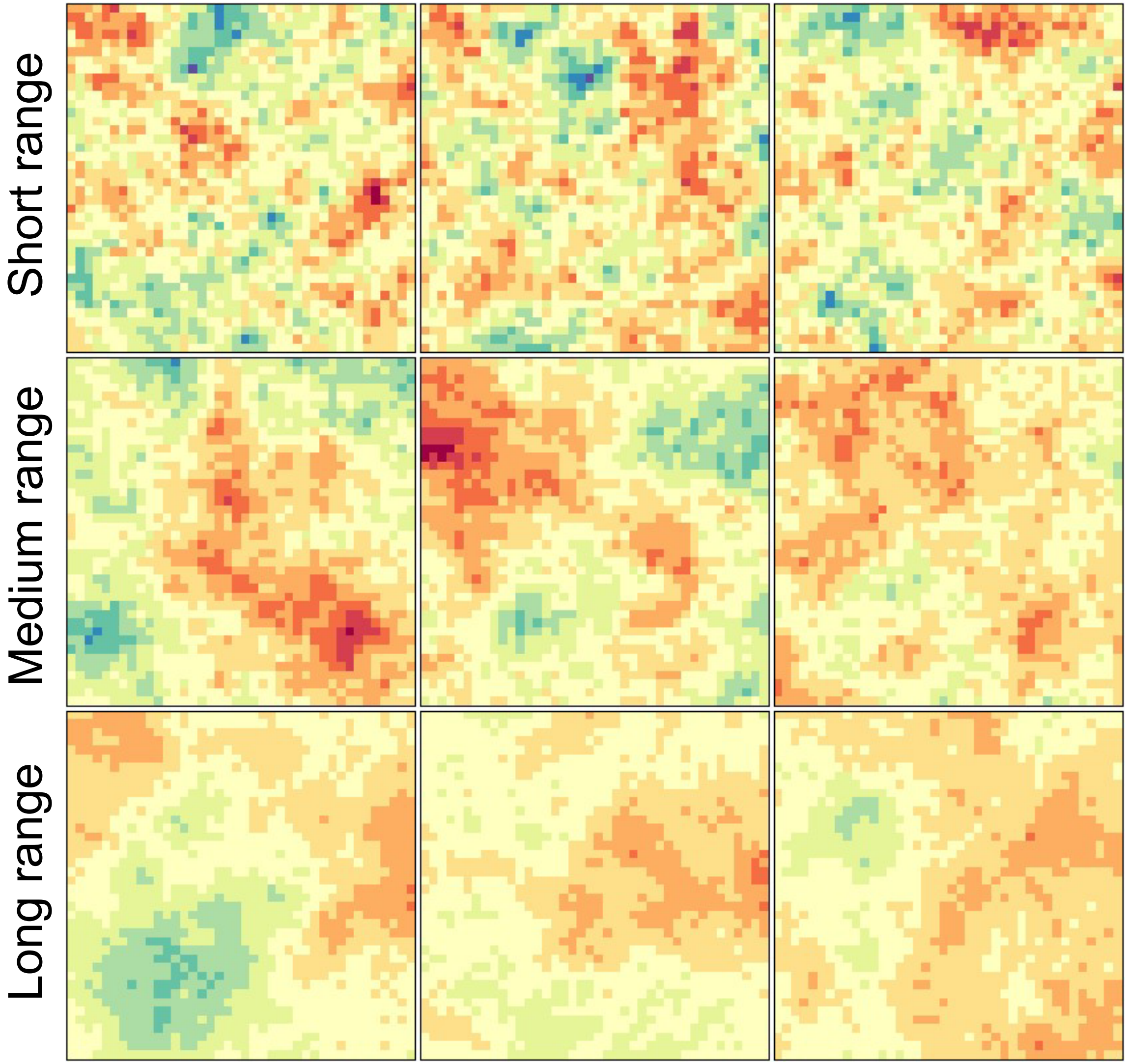
$$\gamma(|s_i - s_j|) = \tau^2 (1 - K(|s_i - s_j|))$$



Examples of Spatial Variation

Example of 9 spatially correlated Gaussian random fields, with each row showing three fields generated with the same spatial correlation structure

An exponential spatial correlation structured was used



Example of Spatial Point Prediction

- **Consider the Model:**

  $Z(s_i) = Y(s_i) + \varepsilon(s_i)$ ← Data

  $Y(s_i) = [\beta_1 + \beta_2 \, lat_i)]$ ← Trend

  $+ \beta_3 x_i$ ← External

  $+ \delta(s_i)$ ← Spatial

  where the unknown parameters are the β's and the variance/covariance parameters associated with $\varepsilon(s_i)$ and $\delta(s_i)$

- **Parameter Estimation:**

  Maximum likelihood (assuming Gaussian distribution of stochastic error terms)

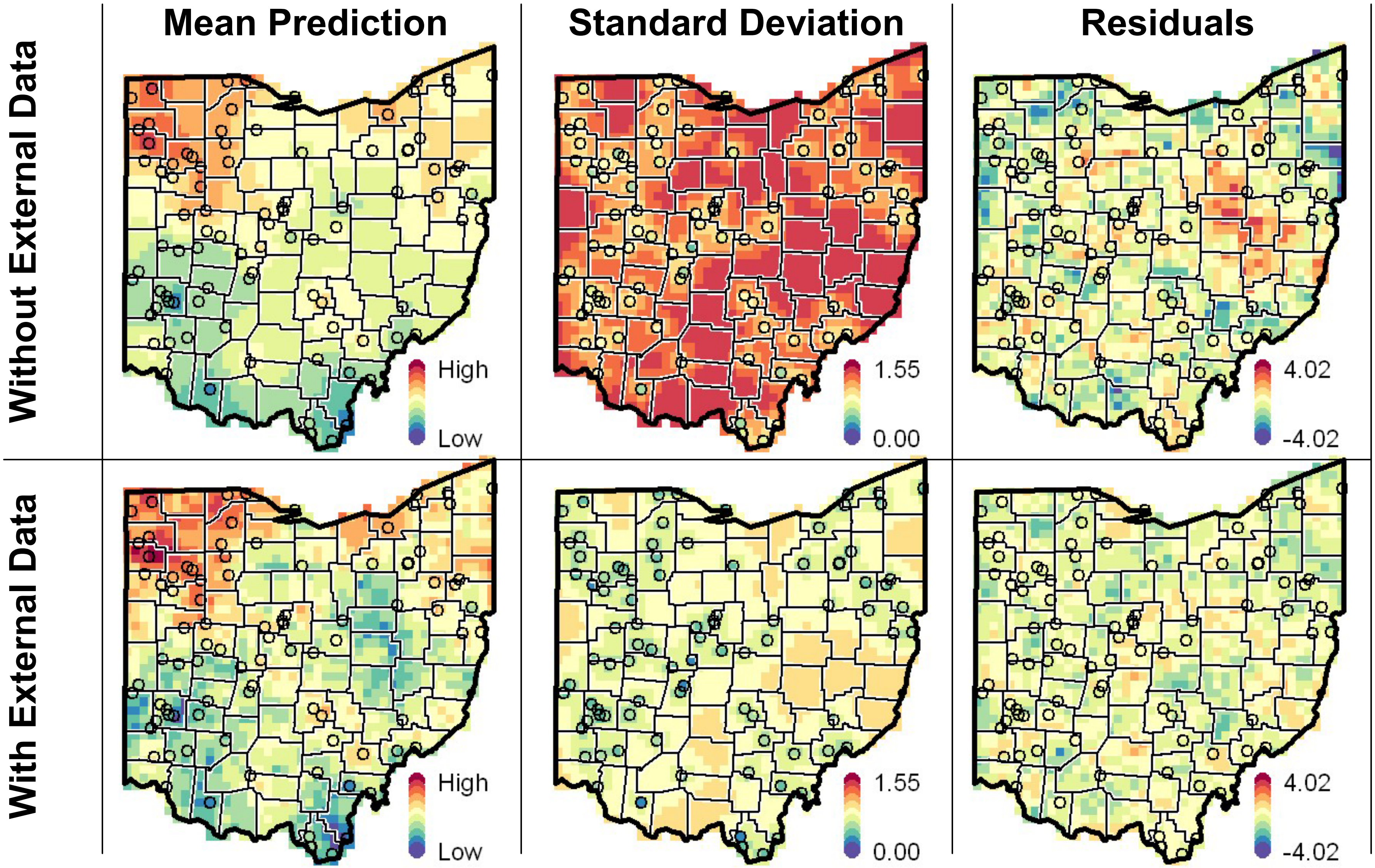
- **Prediction:**

  Given estimates of unknown parameters, the Y-process can be predicted at any spatial location.

  The distribution of the predicted process is *Gaussian*

  Optimal prediction is given by the mean value of the predictive distribution. The standard deviation of the predictive distribution shows the uncertainty

Spatial Point Prediction (con't)



Models for Areal Data

Identical Setup as for Point Data

- **The Data Model**:

  $Z(D_i) = Y(D_i) + \varepsilon(D_i)$

  where:

  $Y(D_i)$ = unknown process

  $\varepsilon(D_i)$ = observation errors

- **The Process Model**:

  $Y(D_i) = \mu(D_i) + f(D_i) + \delta(D_i)$

  where:

  $\mu(D_i)$ = large-scale trend

  $f(D_i)$ = external predictor

  $\delta(D_i)$ = small-scale variation

The difference is in the modeling of the spatially-correlated small-scale variation $\delta(D_i)$

The Small-Scale Process $\delta(D_i)$:

- Spatial correlation is induced through the *proximity matrix*

  $\boldsymbol{W} = (w_{ij})$

- **CAR**: The conditional autoregressive model:

$$\delta_i \mid \delta_{-i} \sim N\left(\frac{\rho}{w_{i+}} \sum_j w_{ij}\delta_j , \frac{\tau^2}{w_{i+}} \right)$$

- **SAR**: The simultaneous autoregressive model:

$$\delta_i = \rho \sum_j w_{ij}\delta_j + u_i$$

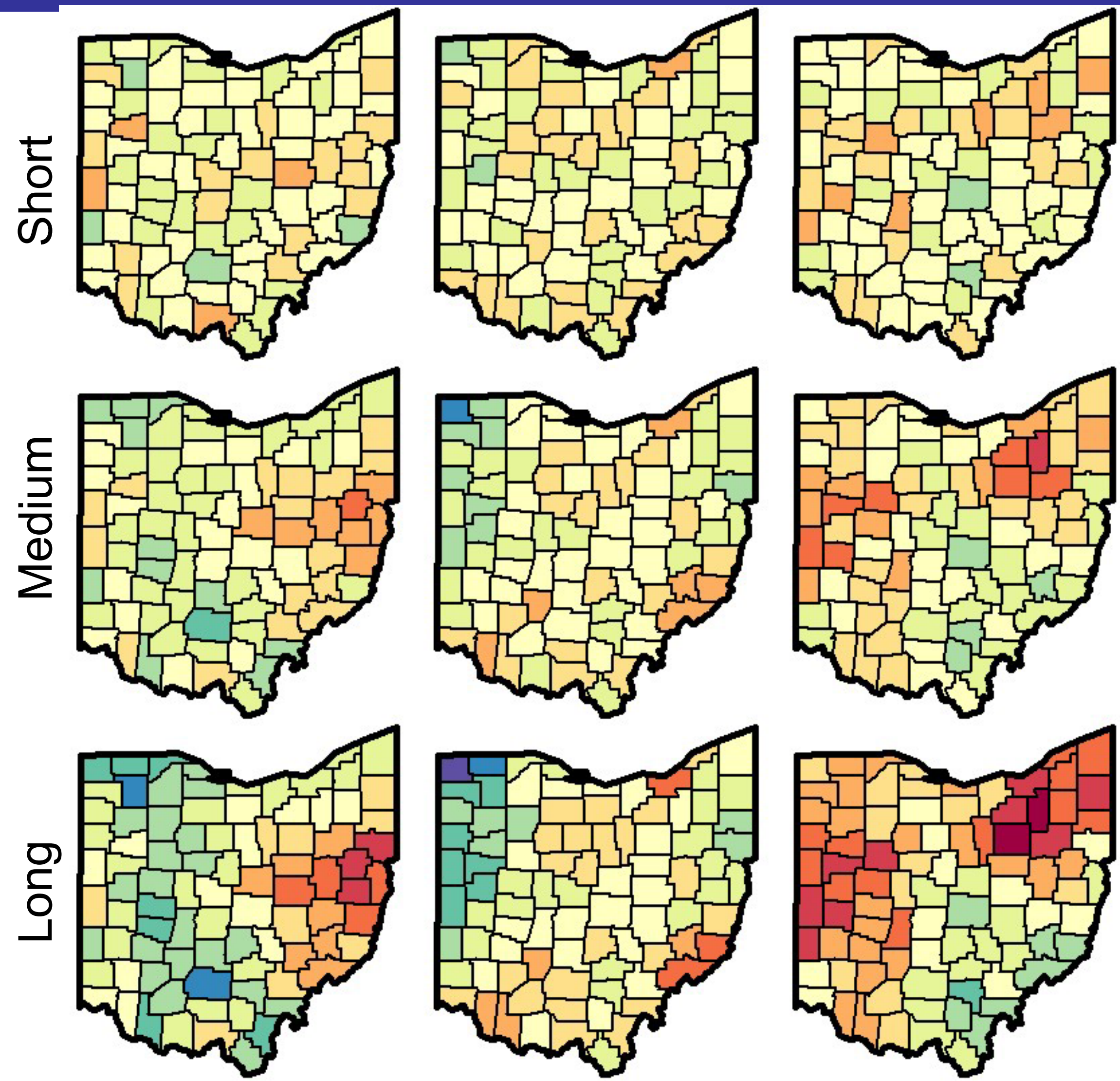
  where $u_i \sim N(0,\tau^2)$

- For "grid-like" areal-units, one can use point-referenced spatial correlation models

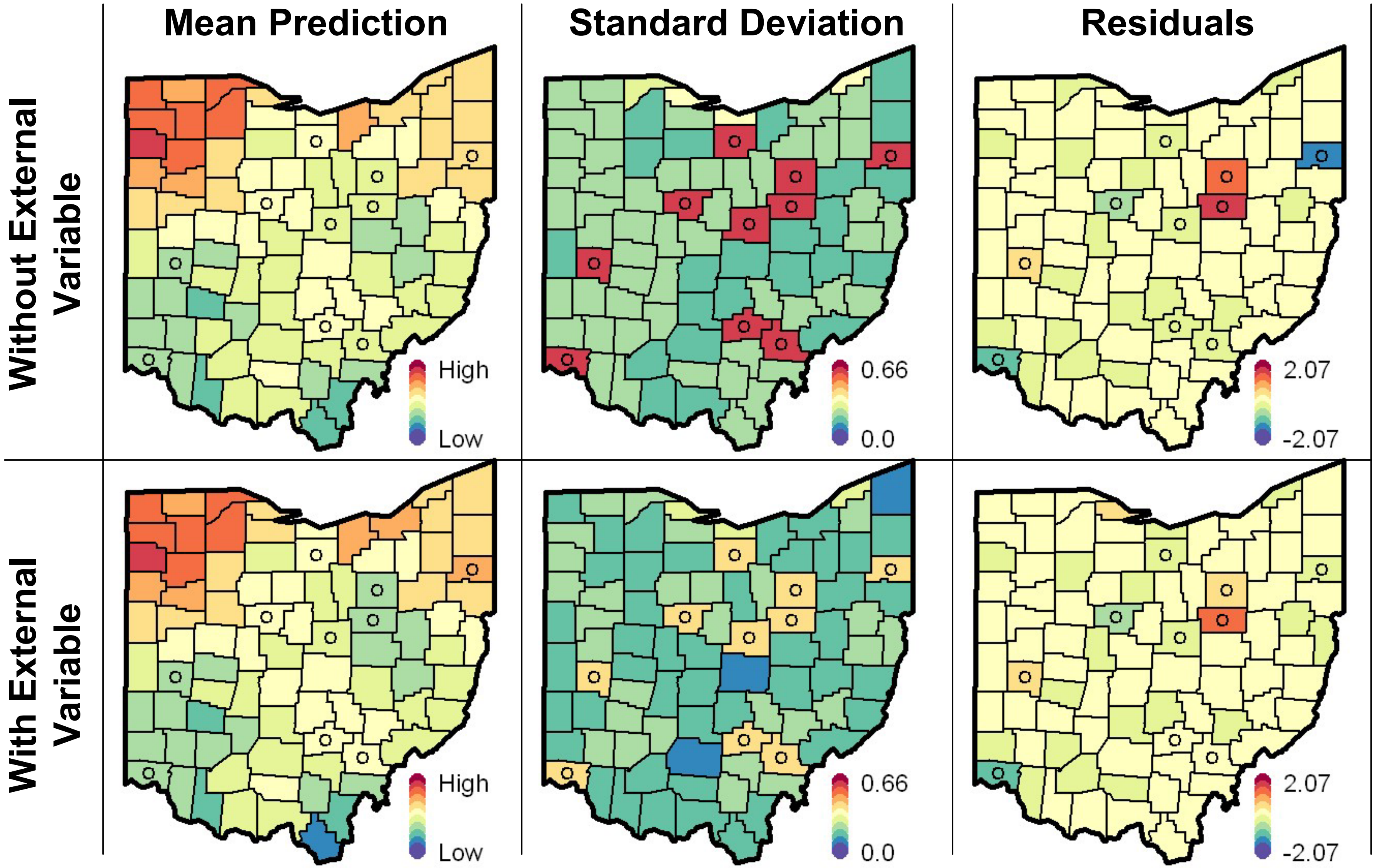
Examples of Spatial Variation

Example of 9 spatially (Gaussian) correlated areal data generated using the SAR model with 3 different spatial correlation ranges.

A nearest-neighbor proximity matrix is used



Example of Areal Unit Prediction



Using Both Point and Areal Data

- It is possible to take simultaneously advantage of point-referenced data and areal data when both carry information about the same process, although at different spatial scales
- Need to model the target process at a point-scale (or at a fine resolution)
- **The Data Model**:

$$Z(s_j) = Y(s_j) + \varepsilon(s_j)$$

$$Z(D_i) = Y(D_i) + \varepsilon(D_i)$$

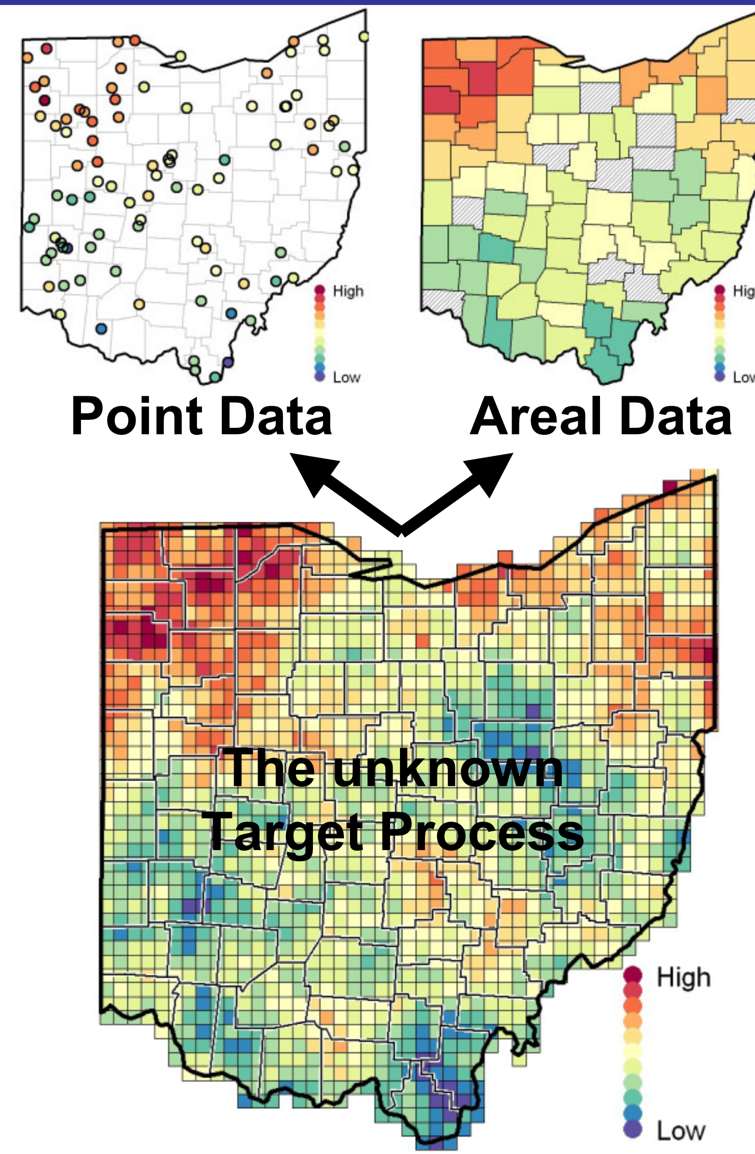
where:

$Y(s)$ = unknown process

$Y(D) = \text{avg}\{Y(s) : s \text{ in } D\}$

$\varepsilon(s_j)$ = point-data error

$\varepsilon(D_i)$ = areal-data error



Using Spatial Prediction Maps

The goal is to use the results of the spatial prediction to answer questions of interest; the plug-in approach vs. the stochastic approach

- **Example of usage**: Want to estimate the average/max/min of the Y-process in a give sub-region
- Recall that the result of the spatial prediction is a *distribution* of the Y-process
- **The plug-in approach**: Use the mean predicted value of the process and compute the quantities of interest; that is, plug the optimal map into the "formula" for the answer
- Uncertainty associated with the plug-in approach: Can use the predicted standard deviation to estimate uncertainty in the output
- **The stochastic approach:** Generate realizations of the Y-process from the predictive distribution (i.e., many maps) and compute the quantities of interest for each map. This yields the *distribution* of the quantities of interest

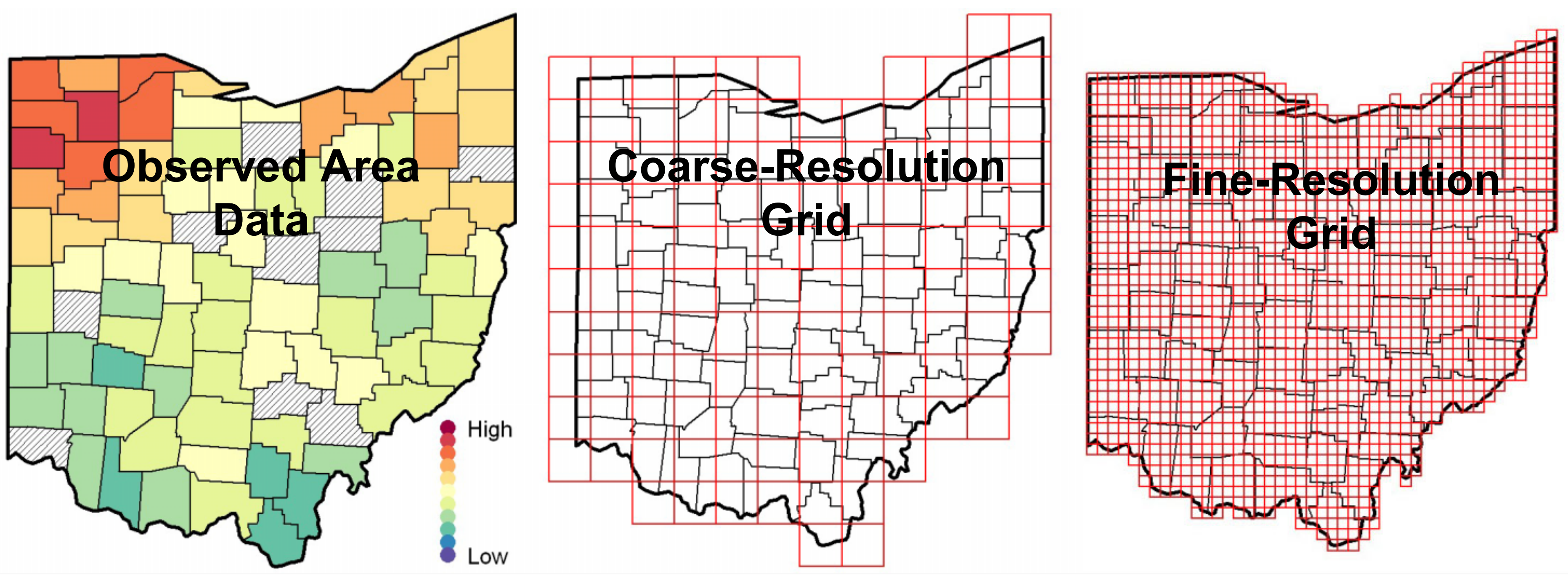
ESDA:Key Take Away Message

- Using ESDA can make better use of Areal and Point data sets than other standard data analysis methods such as Classical Statistics.
- ESDA allows you to build data and prediction models that will provided defensible techniques to data manipulation.

Misaligned Data and Processes

What to do when data is reported at one set of areal units but we need to make inference at a different set of areal units (different scale)

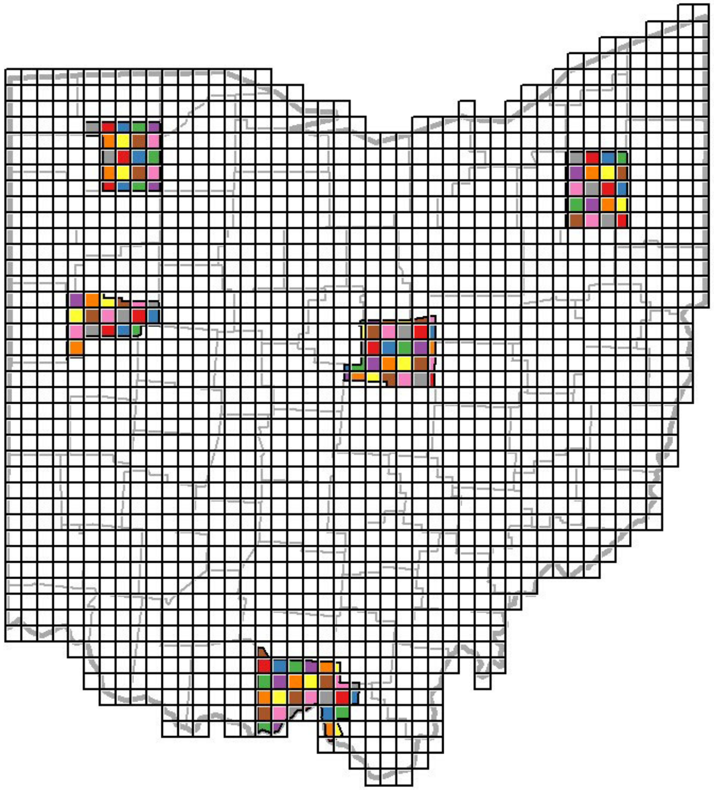
For example: data reported on the county-level (with uncertainty and/or missing data for some counties), but we are interested in the unknown process on a regular coarse or fine resolution grid

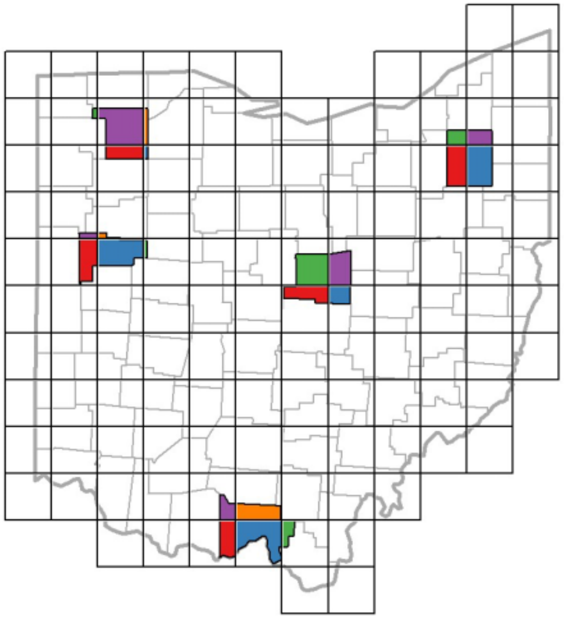


Resampling via GIS Software

A popular approach in GIS software is to "resample" the data to the new areal-units

This is accomplished by assuming that the underlying process is uniformly distributed within each original areal unit and reallocation (resampling) is then carried out in proportion to overlapping area



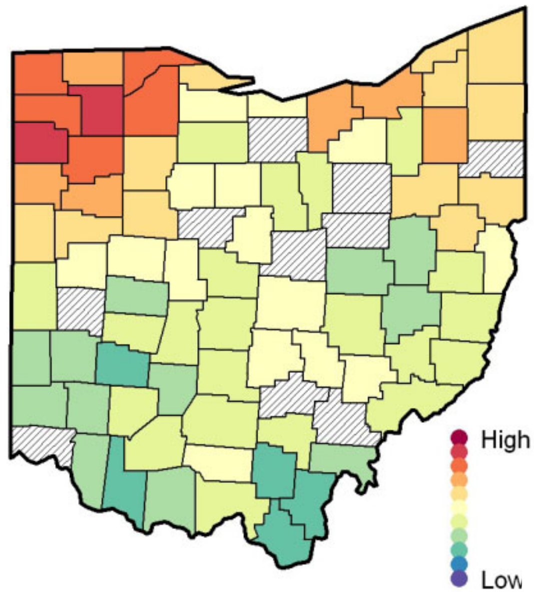


Problems occur if there is

- missing data, or
- the underlying process is not believed to be (approximately) uniform within the original areal units

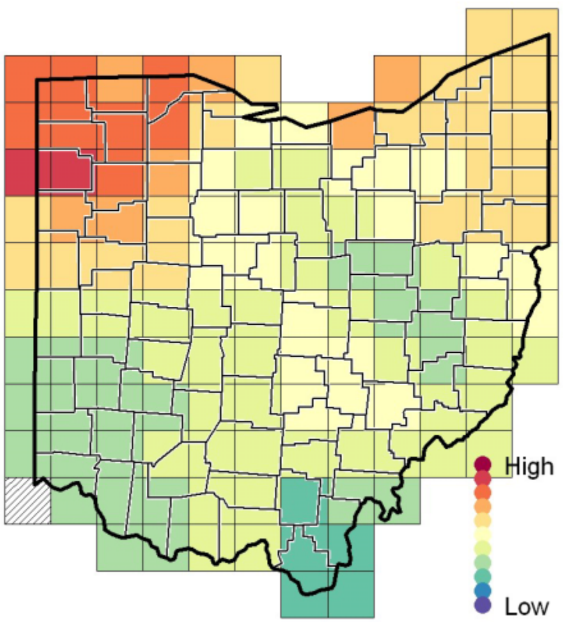
Example of Resampling

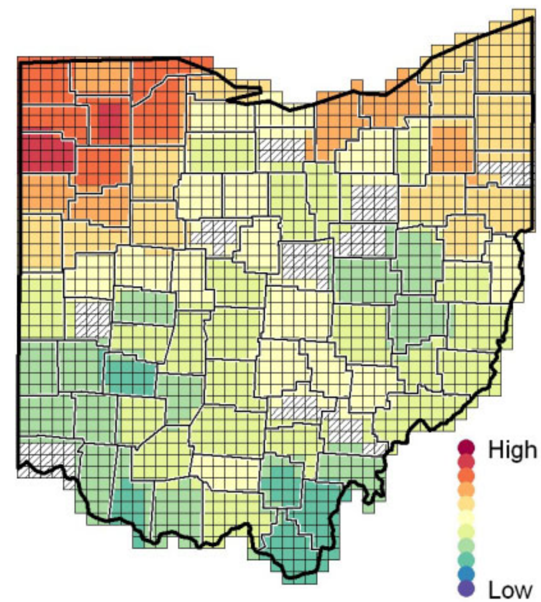
An example of resampling (complete) county-data to a course and fine resolution grids



The results for the fine-resolution grid are not realistic due to how refined it is compare to the original county data

Resampling





Model-Based "Resampling" Approach

The goal is to create a more realistic "resampling" method

The approach taken is to model the unknown Y-process either at:

(1) the target areal units (i.e., fine-resolution grid) or at

(2) the set of areal units formed by the union of the areal units of the data and the target areal units

If the target areal units are at a much finer spatial scale than the data areal units, there is little difference between the two approaches

The **data-model** becomes:

$$Z(D_i) = \sum_j a_{ij} Y(B_j) + \varepsilon(D_i)$$

Where a_{ij} is the area proportion of the data areal unit D_i that is within the fine-resolution cell B_j

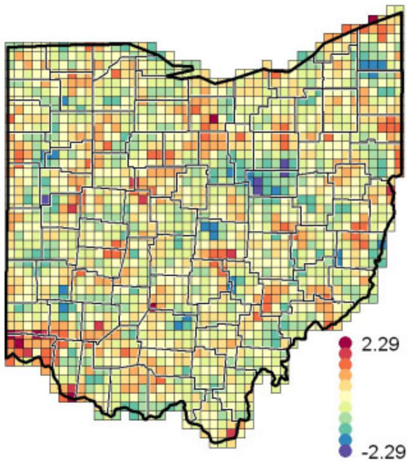
The fine-resolution Y-process is then modeled as an areal process (as introduced previously)

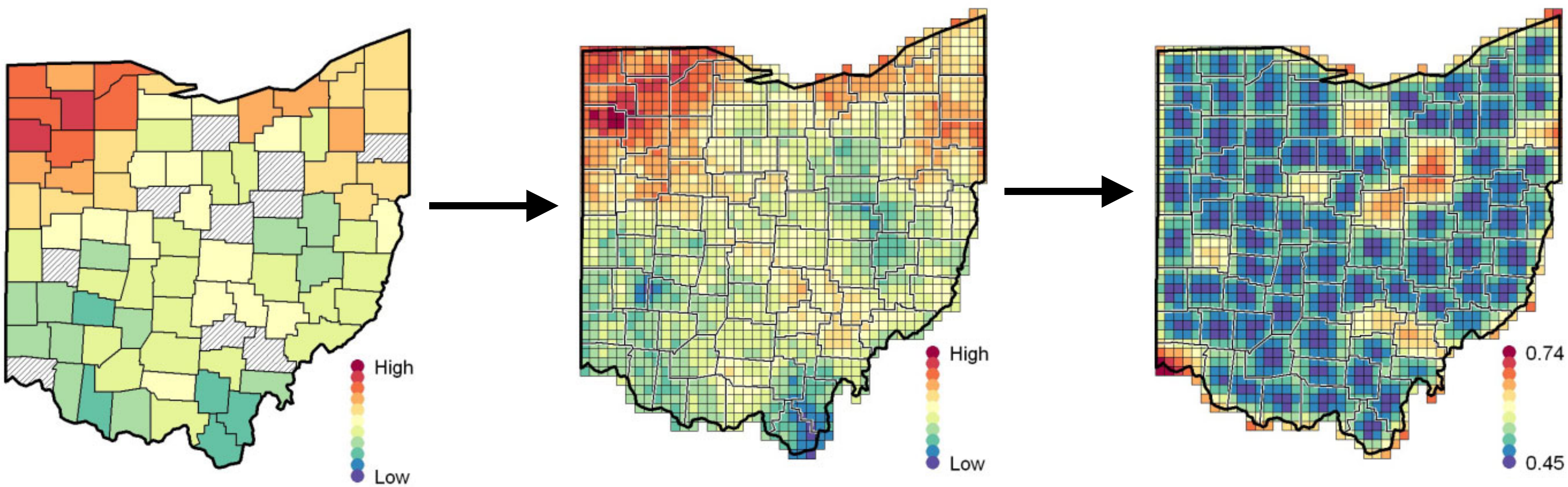
Example of Model-Based Resampling

Given incomplete county-data, the Y-process is predicted at a fine-resolution grid and the associated standard deviations mapped.

Comparison to the true Y-process shows favorable results

The true Y-process minus predicted





Data

Prediction

Standard Deviation

Spatial Representation

What Resolution Should I Use?

We are able to work with multiple data at different spatial scales and predict the unknown Y-process at altogether different spatial scale.

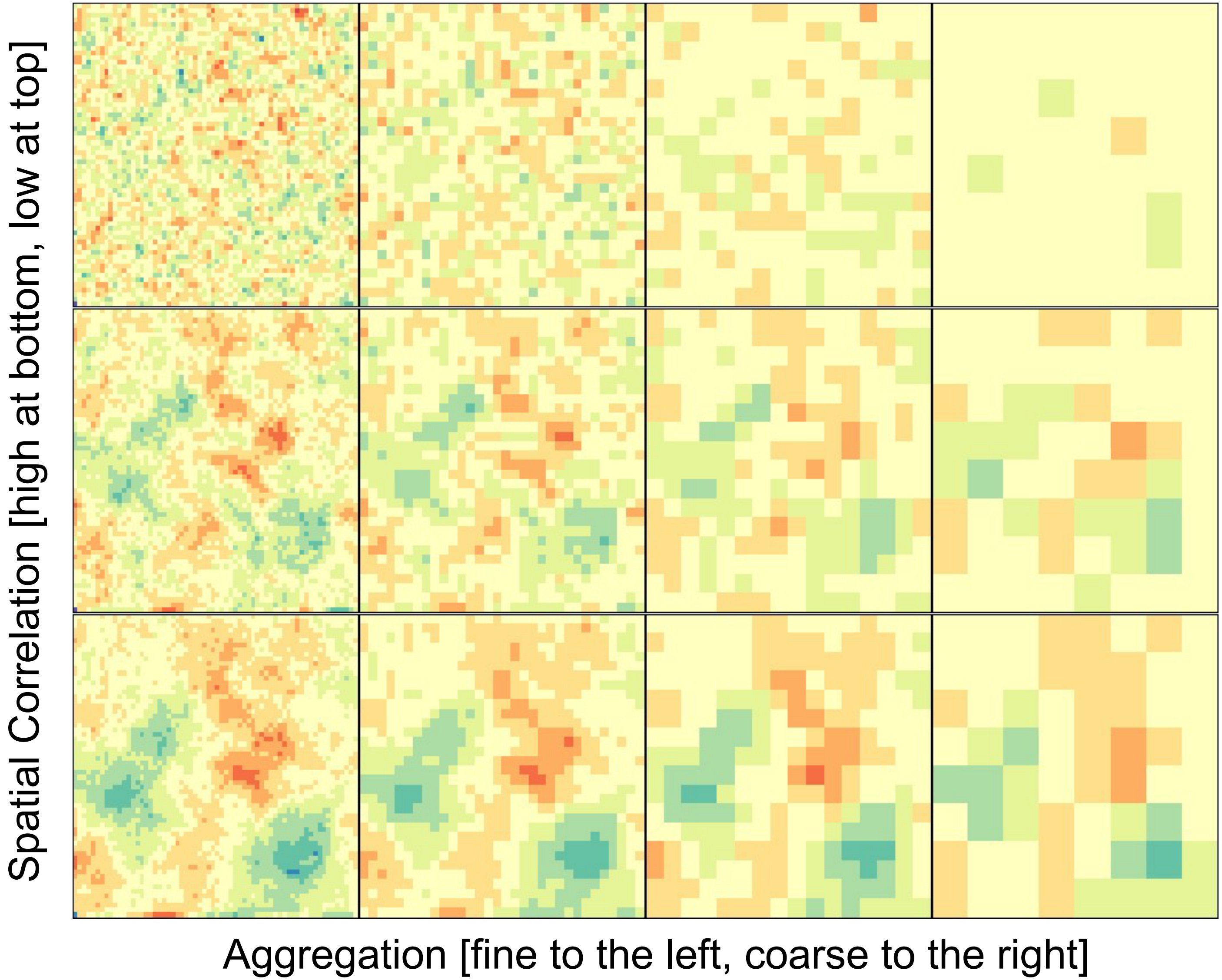
Question: What resolution should I use when predicting the Y-process?

- …It depends on what the prediction map is going to be used for…
  - Aggregation reduces the variation in the underlying process; shrinks the process to the local average.
  - For some applications this might be problematic (e.g., finding the spread of the process), but not so for other applications (e.g., finding the mean).
- …It depends on the (natural) spatial variation in the process…
  - If the process varies slowly with spatial location, due to long-range spatial correlation, then a coarser-resolution representation might be sufficient.

Spatial Correlation and Resolution

Each row of four maps shows the aggregation of a (Gaussian) randomly generated spatial processes

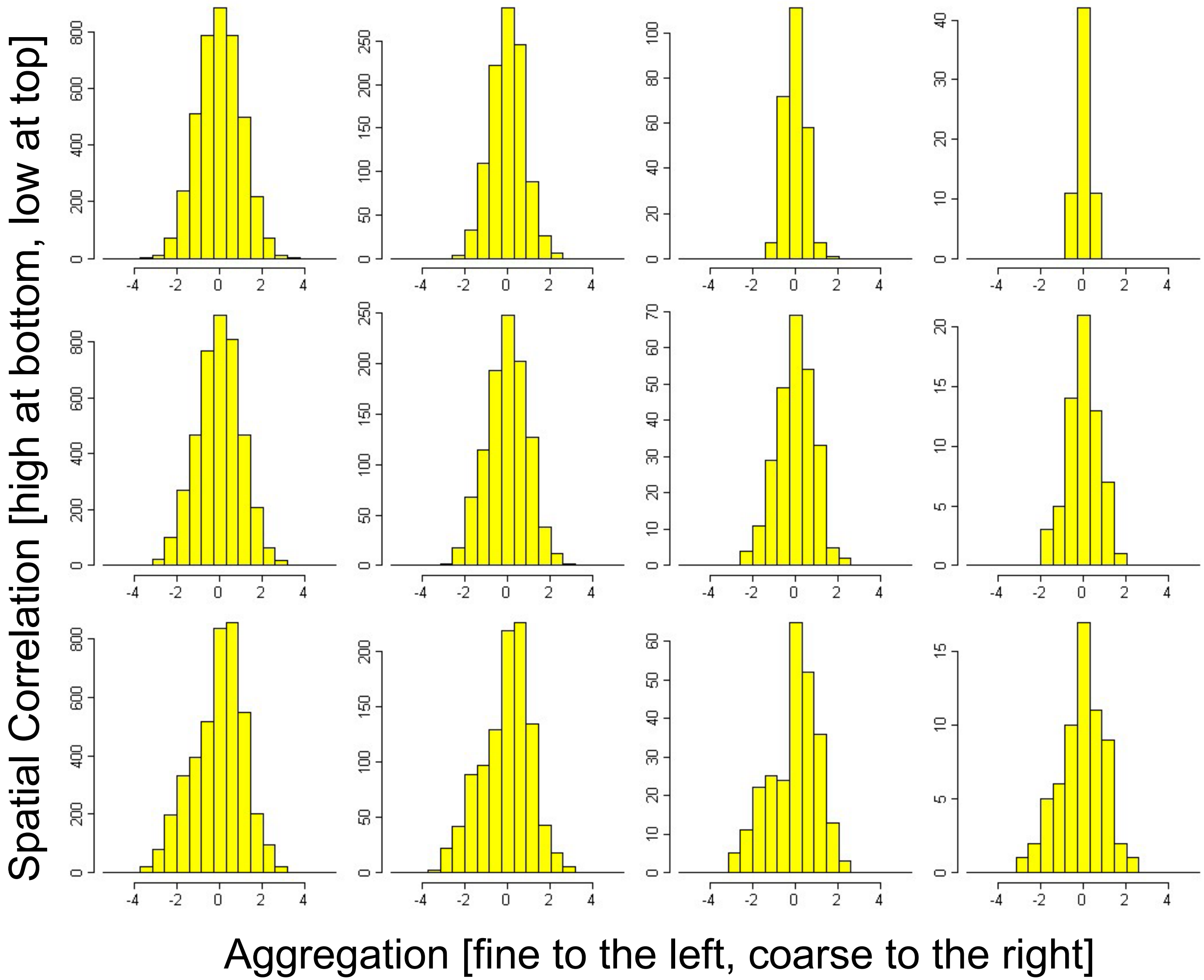
Take note of how the aggregation has lesser impact on the long-range spatially correlated process (bottom) than on the short-range spatially correlated process (top)



The Impact on Variation

Aggregation reduces the variation in the process

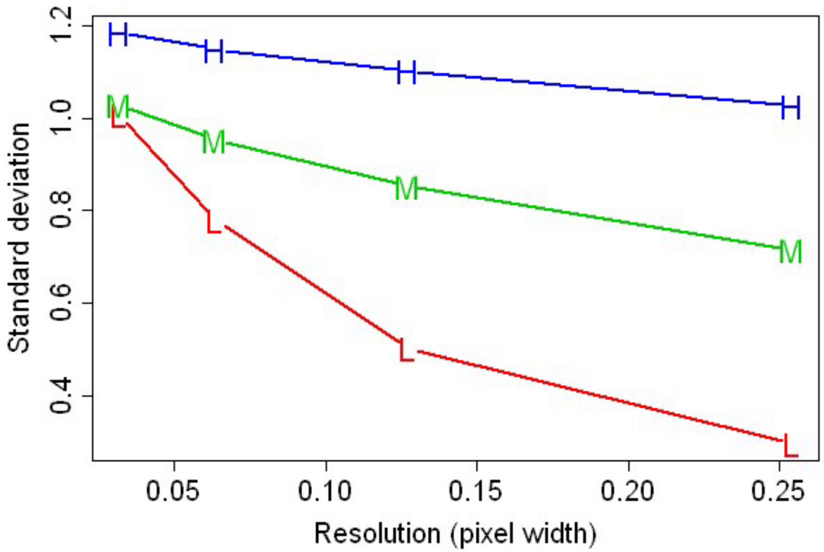
Aggregation has lesser impact on the mean (average) tendency of the process

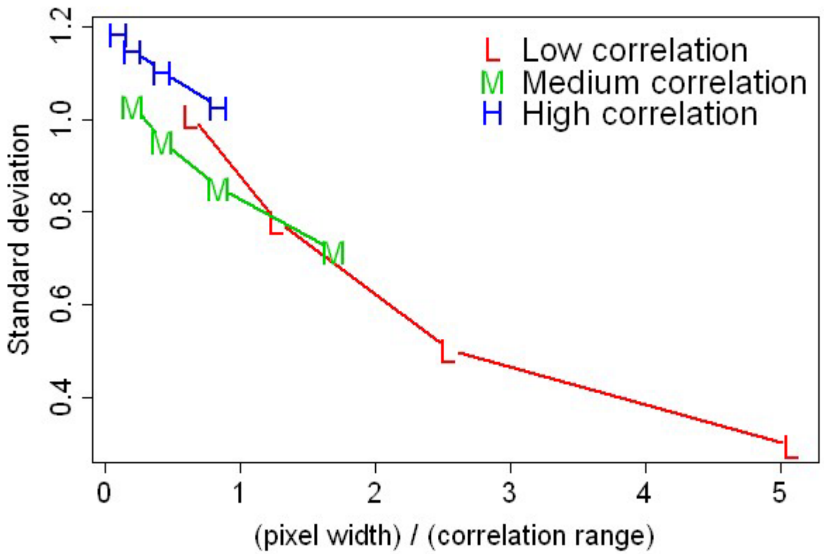


Variation, Correlation, and Resolution

Question: How does the change in variation relate to the spatial correlation and resolution?

- The variation, as represented by the standard deviation, decreases with increasing pixel-width for all ranges of spatial correlation, but more so for processes with low spatial correlation (see top graph),
- The variation, as represented by the standard deviation, decreases with the pixel-with to correlation-range ratio (bottom).
- Given information about the spatial correlation, a resolution (pixel width) can be chosen





(3) Illustrate examples using common EERE data sets.

Geospatial Statistics and Issues in Energy Modeling

Gardar Johannesson and Jeffrey Stewart

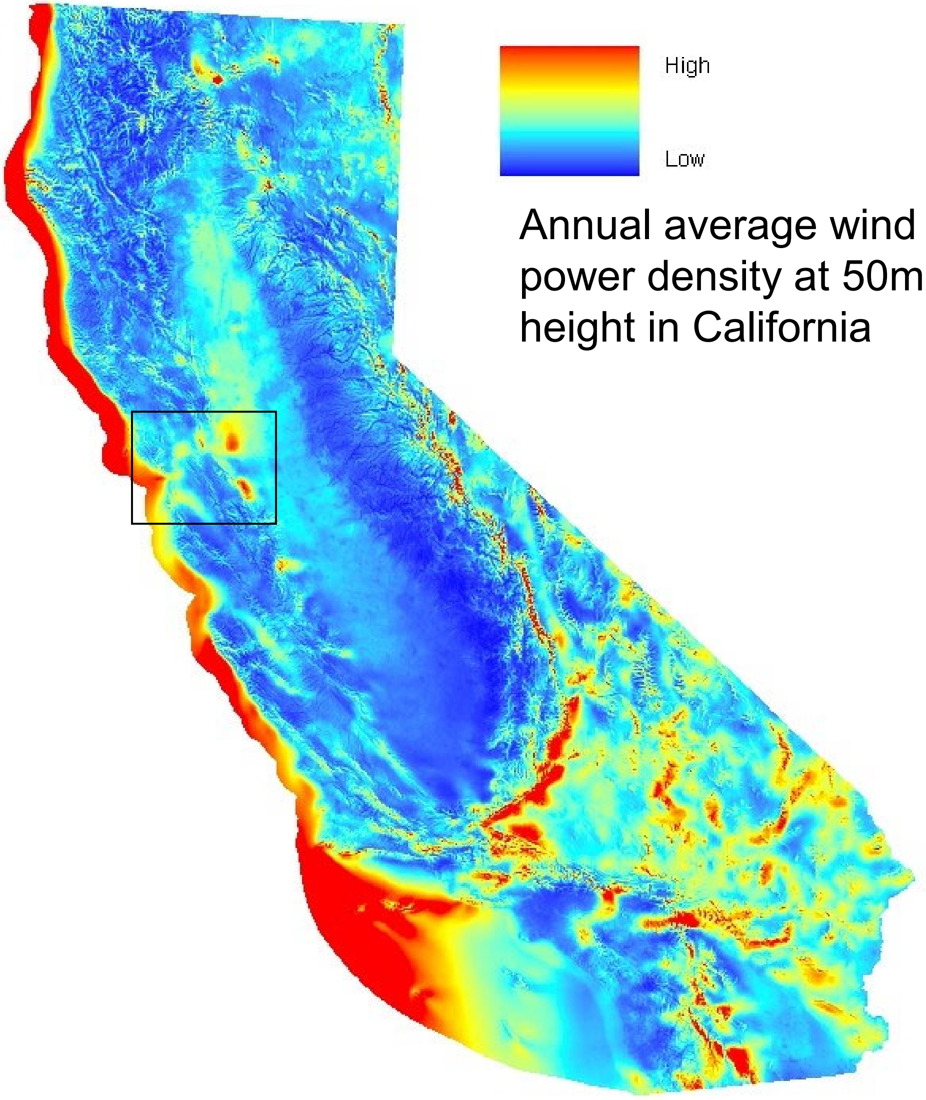
Lawrence Livermore National Laboratory

May 10-11, 2005

Annual Average Wind Power Density Data For California

Annual Average Wind Data

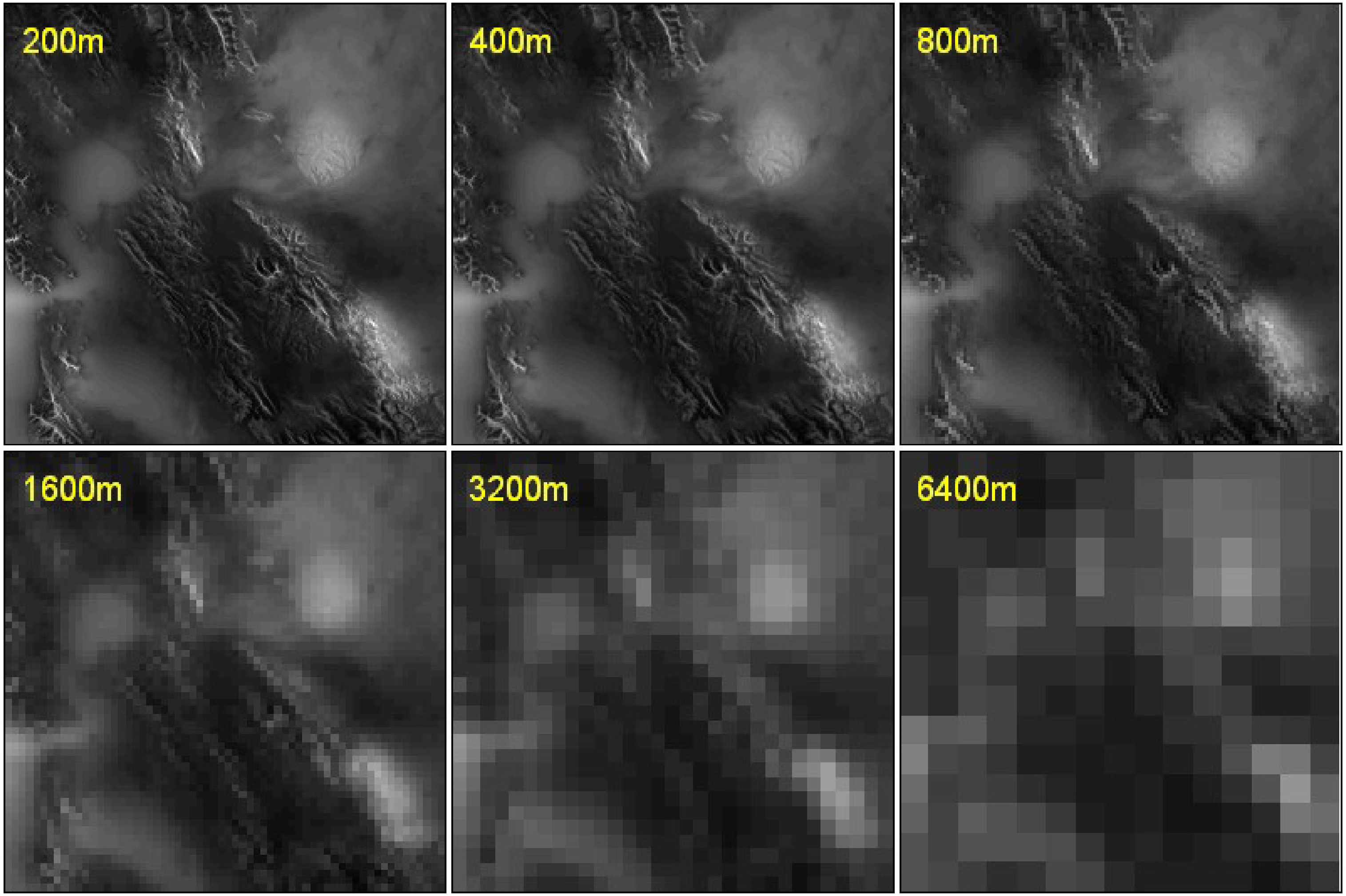
- The data covers California and shows an estimate of the annual average wind power density at 50m height as provided by TrueWind Solutions
- The data is reported as 200x200 meters pixel averages; areal-data
- **The question is**: For geo-processing, is it necessary to work at the 200m resolution or can one work with (more manageable) courser-resolution maps?



Annual average wind power density at 50m height in California

Study Area

A small sub-region was selected for further study and the original 200m resolution data was aggregated to coarser resolutions

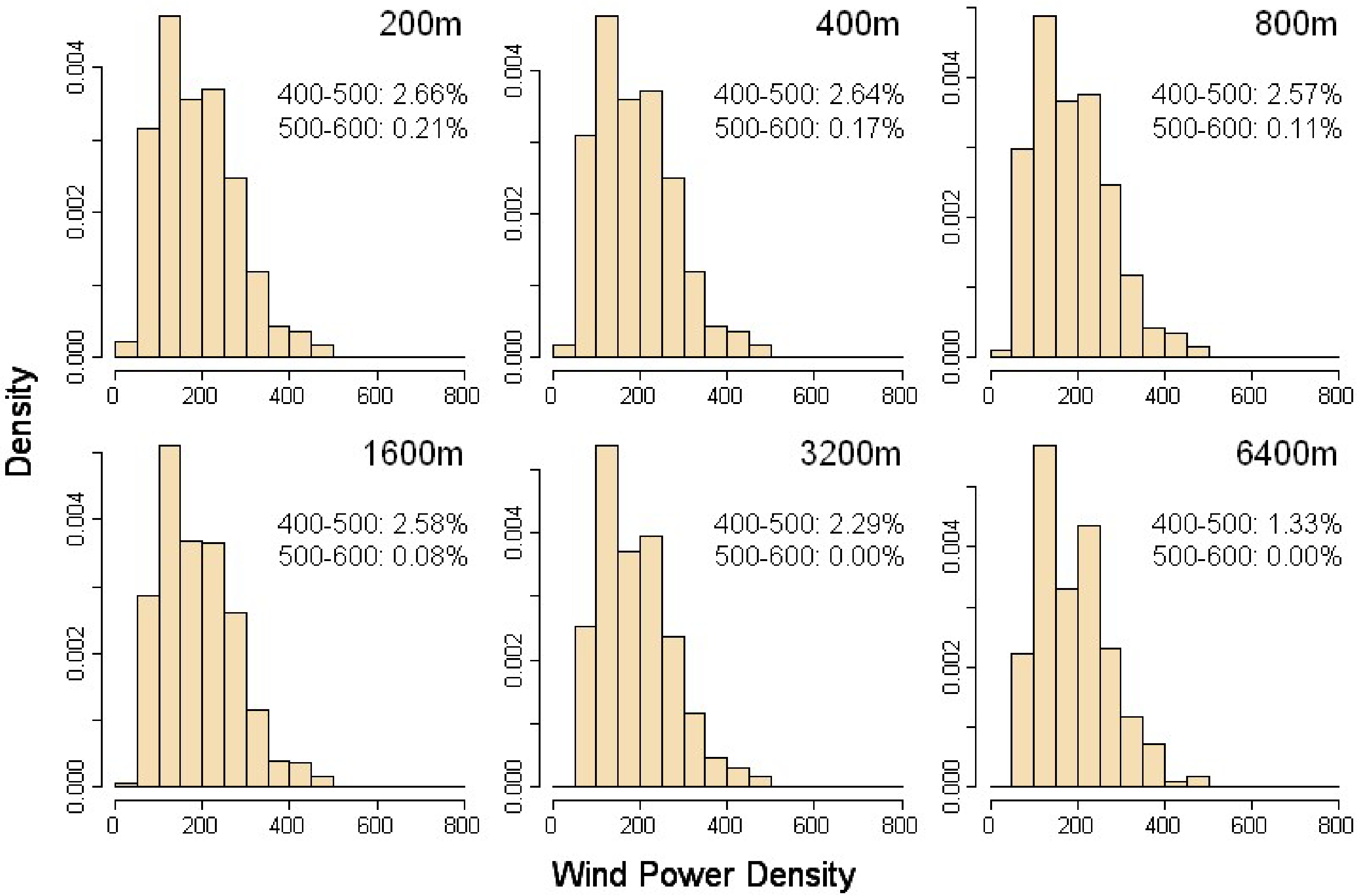


ESDA: Variation

Use histograms to visualize the (marginal) variation in the wind power density at each resolution and look for shrinking variation with resolution

There is very little change in variation with resolution.

The impact on the right-tail of the distribution is important and the fraction of pixels in the range 400-500 (class 4) and in 500-600 (class 5) are shown.

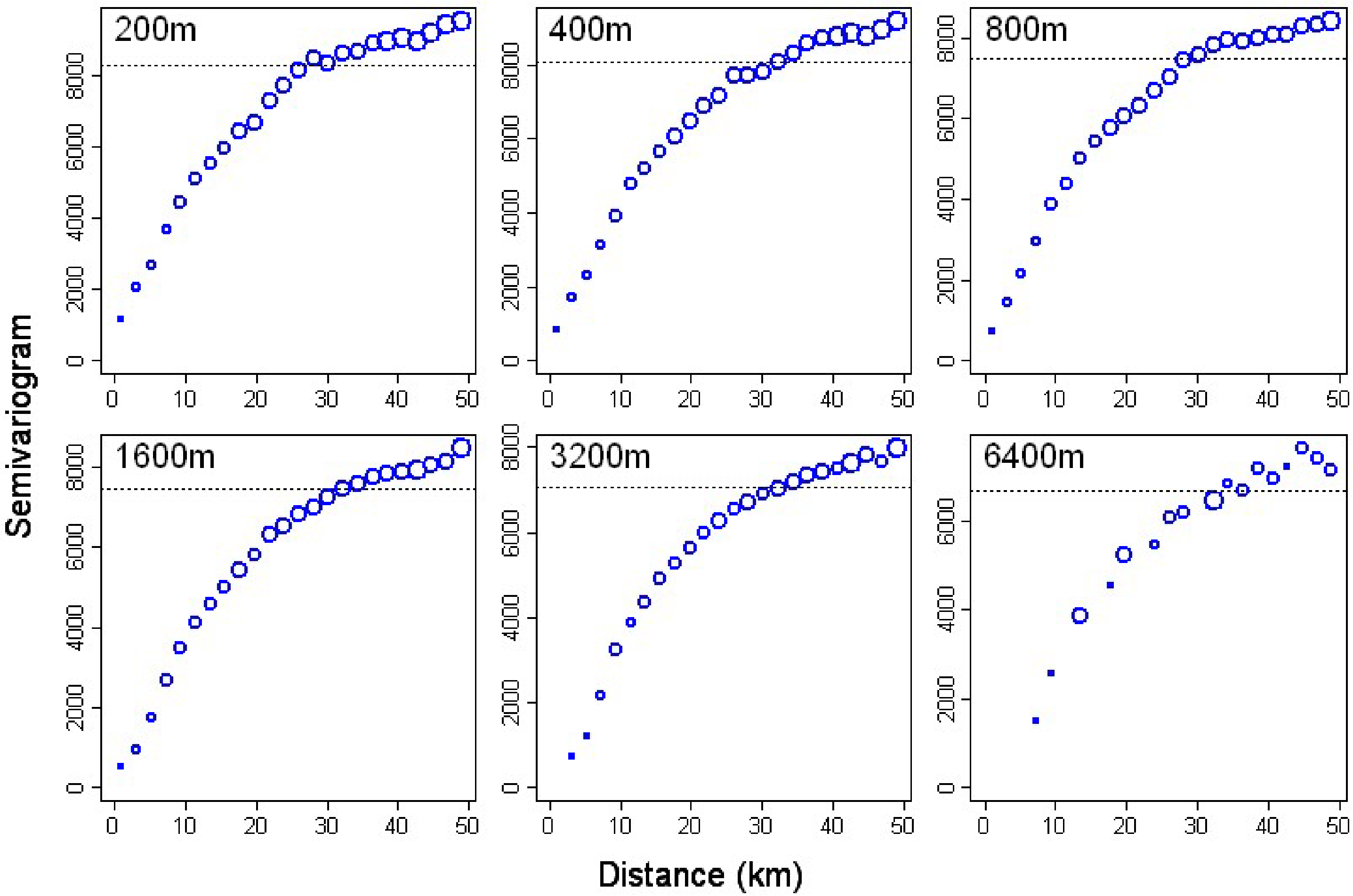


ESDA: Spatial Correlation

The semivariogram can be used to summarize the spatial correlation pattern in the data at each resolution

The spatial correlation is seen to level off around 20km.

Note that the information in the semi-variogram does not change with resolution



ESDA: Conclusion

- The goal was to determine the coarsest resolution that one could work with and still have a good representation of variation in the wind power density, in particular the shape and extend of the right-tail of the distribution.
- To analyze the impact that resolution has on variation, the 200m resolution data was aggregated in five steps down to 6,400m resolution. A histogram of the aggregated wind power densities shows a gradual shrinking of the variation with coarser resolution, particularly when exceeding 3,200m resolution.
- Instead of aggregating to coarser resolutions, it is possible to use the semivariogram, derived from the original 200m resolution data, to infer about the amount of spatial correlation in the data. The spatial-correlation range provides information about the impact that aggregation at different resolutions has on the variation in the data.